CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Enhancing Author Name Disambiguation

by

Humaira Liaquat

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

in the

Faculty of Computing
Department of Computer Science

2024

Enhancing Author Name Disambiguation

By

Humaira Liaquat

(DCS163003)

Dr. Donghong Ji, Professor

Wuhan University, Wuhan, Hubei, China

(Foreign Evaluator 1)

Dr. Hassan Malik, Senior Lecturer

University of Essex, Essex, UK

(Foreign Evaluator 2)

Dr. Muhammad Abdul Qadir

(Research Supervisor)

Dr. Abdul Basit Siddiqui

(Head, Department of Computer Science)

Dr. Muhammad Abdul Qadir

(Dean, Faculty of Computing)

DEPARTMENT OF COMPUTER SCIENCE

CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY

ISLAMABAD

2024

To my family, especially my father.

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

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in the dissertation, entitled "**Enhancing Author Name Disambiguation**" was conducted under the supervision of **Dr. Muhammad Abdul Qadir**. No part of this dissertation has been submitted anywhere else for any other degree. This dissertation is submitted to the **Department of Computer Science, Capital University of Science and Technology** in partial fulfillment of the requirements for the degree of Doctor in Philosophy in the field of **Computer Science.** The open defence of the dissertation was conducted on **August 05, 2024**.

Student Name : Humaira Liaquat (DCS163003)

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

Examination Committee :

(a) External Examiner 1: Dr. Rabeeh Ayaz Abbasi,
Professor
QAU, Islamabad

(b) External Examiner 2: Dr. Amanullah Yasin,
Associate Professor
Air University, Islamabad

(c) Internal Examiner : Dr. Nadeem Anjum
Professor
CUST, Islamabad

Supervisor Name : Dr. Muhammad Abdul Qadir
Professor
CUST, Islamabad

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

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

AUTHOR'S DECLARATION

I, **Humaira Liaquat** (**Registration No. DCS163003**), hereby state that my dissertation titled, '**Enhancing Author Name Disambiguation'** is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/ world.

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

(**Humaira Liaquat**)

Dated: 05, August, 2024

Registration No: DCS163003

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the dissertation titled "**Enhancing Author Name Disambiguation**" is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete dissertation has been written by me.

I understand the zero-tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled dissertation declare that no portion of my dissertation has been plagiarized and any material used as reference is properly referred/ cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled dissertation even after award of PhD Degree, the University reserves the right to withdraw/ revoke my PhD degree and that HEC and the University have the right to publish my name on the HEC/ University Website on which names of students are placed who submitted plagiarized dissertation.

(**Humaira Liaquat**)

Dated: 05, August, 2024

Registration No: DCS163003

List of Publications

It is certified that the following publication(s) have been made out of the research work that has been carried out for this dissertation:-

1. **H. Waqas**, M. A. Qadir, "Completing features for author name disambiguation (AND): an empirical analysis," *Scientometrics* , vol. 127.2, pp. 1039-1063, 2022.

2. **H. Waqas**, M. A. Qadir, "Multilayer heuristics based clustering framework (MHCF) for author name disambiguation," *Scientometrics* , vol. 126.9, pp. 7637-7678, 2021.

(Humaira Liaquat)
Registration No: DCS163003

Acknowledgement

"If you are grateful, I will surely increase you in favor"

Surah Ibrahim 14:7.

I express my gratitude to the Almighty Allah for granting me the strength, courage, and resilience to successfully complete this endeavor.

After this, I dedicate this dissertation to the remarkable individuals who have played a significant role in my academic journey and have been instrumental in my success.

To my esteemed supervisor Professor, Dr. Muhammad Abdul Qadir, I am deeply grateful for your guidance, wisdom, and unwavering support throughout my research. Your expertise, encouragement, and valuable insights have shaped my understanding and enhanced my scholarly abilities. I am truly fortunate to have had the opportunity to work under your mentorship.

To my loving husband, Group Captain Waqas Tahir, your constant support, understanding, and belief in me have been my pillar of strength during the challenging times of this journey. Your unwavering love and encouragement have fueled my determination to overcome obstacles and pursue excellence. I am forever grateful for your presence in my life.

To my wonderful children, Aamnah and Ahmed, you have been my greatest motivation and source of inspiration. Your endless love, understanding, and patience have kept me going even when faced with the toughest of challenges. Your innocent smiles and unwavering belief in me have given me the strength to persevere and achieve my goals.

Above all, to my parents, especially my father, your unwavering faith in my abilities and your constant reminder of the importance of hard work, dedication, and

never giving up on my dreams have been the driving force behind my accomplishments. Your sacrifices, guidance, and unconditional love have shaped me into the person I am today. I am forever indebted to you for instilling in me the values of perseverance and determination.

This dissertation is a testament to the immense support and love I have received from each one of you. Your presence in my life has made this journey not only academically fulfilling but also personally enriching. I am eternally grateful for your unwavering belief in me and for being my pillars of strength.

With heartfelt gratitude,

(Humaira Liaquat)

Abstract

Author name ambiguity problem arises when multiple authors have identical names or when variations of an author's name resemble those of other researchers. This issue affects the accuracy of academic authorship in digital libraries and scholarly data search engines. Despite substantial efforts to address this problem, it remains a persistent challenge. Existing author name disambiguation techniques measure their performance using precision, recall, and F1 scores. While recall is not a major concern, achieving high recall often results in low precision and vice versa, thereby reducing the F1 scores. It is observed that majority of the existing techniques F1 scores fall in the range of 66-77%, which needs improvements. To enhance the results of such techniques, typically two factors play an important role: the learner model used and the input data provided to it. Improving predictive accuracy in a model is contingent on the provision of relevant, independent, and useful features. However, studies that have delved into this area lack a comprehensive analysis of feature ranking and combinations that enhance the results. To conduct this analysis, a review of existing publicly available datasets is performed, emphasizing the availability of maximum features in them. Nonetheless, it is observed that majority of them provide limited feature coverage. Additionally, they are tailored for specific domains and contain author names predominantly from specific ethnic backgrounds, that too unevenly distributed. Consequently, the applicability of such datasets is limited when aiming to develop solutions that can be generalized across diverse areas, scenarios, and contexts. This research addresses these gaps with the following contributions. First, feature ranking and identification of feature combinations is performed. Second, these features are used by the proposed heuristics-based author name disambiguation approach (MHCF), along with the proposed Research2Vec model to enhance author name disambiguation. MHCF is compared with nine techniques in total, using two datasets, i.e., BDBComp and Arnetminer. MHCF comparison (using BDBComp dataset) with SAND1, SAND2, HHC, MHCF-G, and MHCF-GL show a visible gain of 31%, 22%, 32%, 4%, and 1% in enhancing F1 scores. Similarly, MHCF comparison (using Arnetminer) with MDC, GFAD, ATGEP, ESMD, MHCF-G, and MHCF-GL show a visible gain in

F1 scores as 12%, 18%, 32%, 3%, 5% and 10% respectively. Lastly, "CustAND" dataset is proposed to fill the gaps in publicly available datasets, where it is utilized in the feature analysis study, and also by MHCF. This study holds broad implications with wide-ranging applications. Notably, the MHCF methodology offers potential benefits for digital libraries and scholarly search engines. This will facilitate reliable bibliometric calculations, unveiling the genuine scholarly impact and research productivity of researchers. Additionally, it aids various organizations and professionals by identifying potential collaborators, thereby promoting interdisciplinary research, funding opportunities, and various research prospects.

Contents

List of Figures

List of Tables

Abbreviations

AND	Author Name Disambiguation
ACP	Average Cluster Purity
AAP	Average Author Purity
CF1	Cluster F1
CP	Cluster Precision
CR	Cluster Recall
CS	Computer Science
DS	Design Science
DSR	Design Science Research
DLs	Digital Libraries
EG	Ethnic Group
EGs	Ethnic Groups
F1	Geometric mean of precision and recall
GS	Google Scholar
HEC	Higher Education Commission
IT	Information Technology
IOT	Internet Of Things
LSA	Latent Semantic Analysis
MAKG	Microsoft Academic Knowledge Graph
MHCF	Multilayer Heuristics Based Framework
NB	Naive Bayes
NNLM	Neural Network Language Model
P	Precision

pF1	Pairwise F1
pP	Pairwise Precision
pR	Pairwise Recall
PCA	Principal Component Analysis
R	Recall
RF	Random Forest
RF-P	Randon Forest with Permutation
RegEx	Regular Expression
RG	ResearchGate
SFS	Sequential Forward Feature Selection
SVM	Support Vector Machine
TFIDF	Term Frequency Inverse Document Frequency
WoS	Web of Science

Symbols

b^3	performance metrics to evaluate the accuracy of the results in information retrieval
κ	Cohen's kappa co-efficient
w	Weight
$\triangle$	Change
$\rightarrow$	The statement to the left of the arrow implies or leads to the statement on the right
$\leftarrow$	Represent the assignment or initialization operation
$\sum$	Loop
$\surd$	Square root
%	Percentange
$*$	Multiply
/	Divide
#	Number
$\geq$	Greater than or equal to
$>$	Greater than
$\leq$	Less than or equal to
$<$	Less than
$\in$	Belongs to
$=$	Equals

Chapter 1

Introduction

1.1 Background

Scholarly data search engines index and provide information regarding authors and their publications. Users use this information to gauge or asses the researcher's contributions in the form of looking at the publications detail, citations counts, h-index, and more. These systems typically use automatic crawlers to crawl different sources to populate the data in their large data collection and associate the authors with their respective publications. This process is highly susceptible to errors, as different authors have similarities in their names [1]. Repercussions of inclusion of these errors are witnessed in the form of inaccurate citation counts, h-index, and overall author's rankings which makes the researcher's contribution doubtful.

To further elaborate the problem and its implication, Figure 1.1 shows an author's profile[1] on Google Scholar (GS). Out of first twenty publications (listed on the first page), twelve are not correctly associated with the author. The citations count of the first twenty publications is 1496. Among which, 1198 citations belong to publications that are incorrectly associated with the author, which makes the resulting precision to merely 19.9%. In the same figure, on careful analysis of

[1]The results shown are considered at the time of first draft of this writing i.e. before 16^{th} December 2022

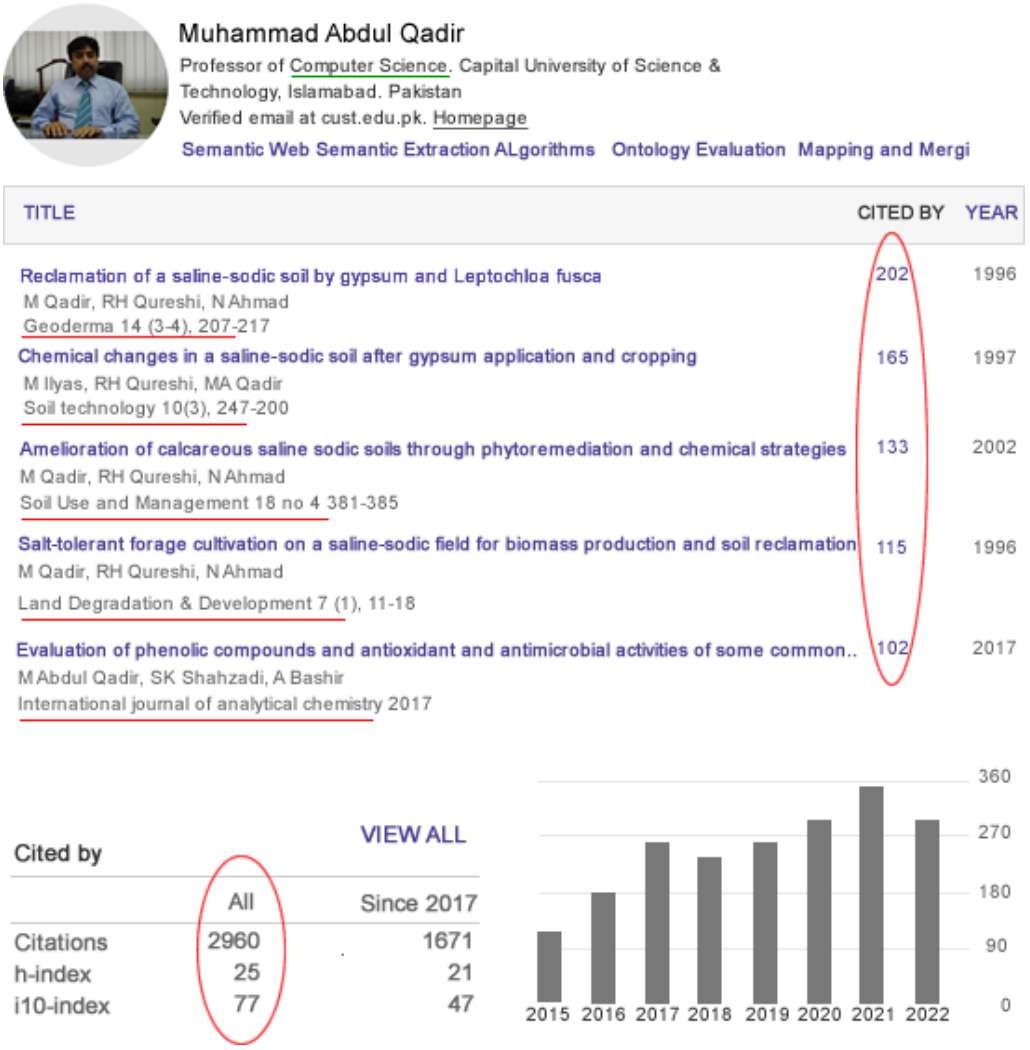

FIGURE 1.1: Example of incorrect academic authorship in Google Scholar.

accumulated citations count, it can be seen that out of 2960 at least 1198 citations are of publications that are wrongly associated with the author. Similarly, out of 1671 citations count of publications since 2017, at least 247 belong to publications authored by someone who has a similar name to the author under discussion. (Note: all the statistics shown and discussed are gathered before 16^{th} December 2022).

These incorrect author publication linkages affect the overall decisions of scientific organizations, policy makers, universities and research agencies for making critical decisions regarding hiring, promotions, funding, etc. It, therefore, becomes very important to disambiguate the authors so that problems occurring because of author name ambiguity can be alleviated [2, 3]. In literature this problem is called

author name ambiguity and removal of incorrect associations of papers is called author name disambiguation (AND).

1.2 Application Scenarios of Author Name Disambiguation

Owing to the continuous increase of scientific and technological publications, there is a growing need for systems that can automatically assign academic authorships. Consequently, the author name ambiguity problem has become an inevitable problem in scholarly search engines, digital libraries [1], and knowledge graphs [4]. To understand the potential application scenarios of AND techniques, consider figures 1.2 and 1.3 which show the basic components of academic search engines and digital libraries. These figures identify the potential components in which these systems can be used, as explained under:

1. **Academic Search Engine**: The primary components of any academic search engine include the crawler, metadata extractors, indexers, and searching components which are responsible for crawling the web to discover and collect academic content, content parsing, metadata extractions, indexing and storing the collected data in a structured format for efficient and accurate retrieval, also identified by S.M. Kumar [5] et al. A crucial aspect of crawling, metadata extraction, indexing, and searching processes is the precise association of authors to their publications. Which can enable:

   (a) precise and relevant search results.

   (b) facilitates information retrieval for researchers and scholars.

   (c) enhancement in the search process, the overall user experience, and their confidence in the validity of search engine results.

2. **Digital Libraries**: In digital libraries, the potential components that can use AND techniques are somewhat similar to academic search engines like

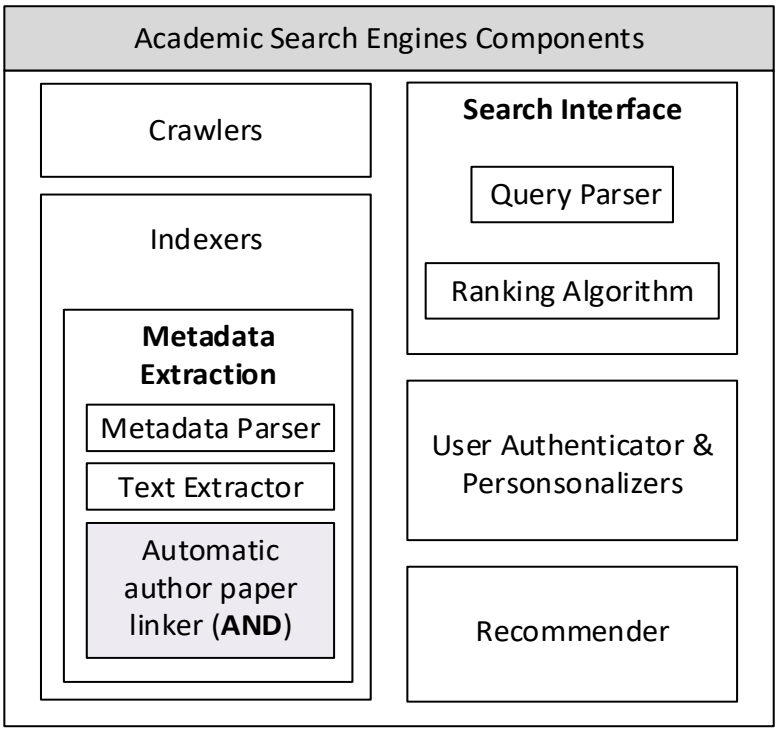

FIGURE 1.2: Application of Automatic Academic Authorship (AND) System in Academic Search Engines.

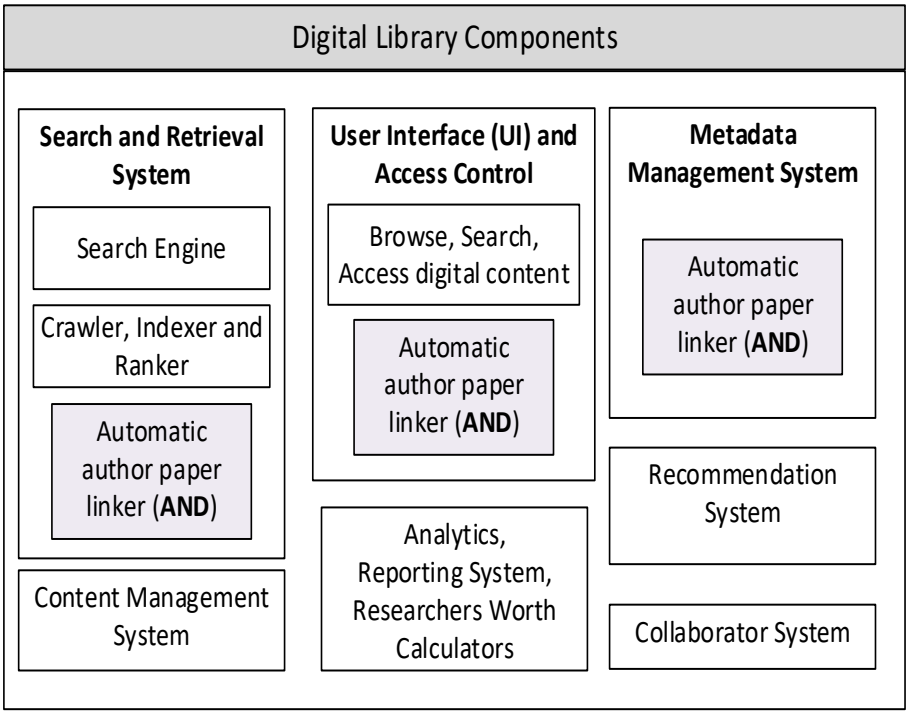

FIGURE 1.3: Application of Automatic Academic Authorship (AND) System in Digital Libraries.

indexers, metadata management systems, and searching systems. The architecture of the digital libraries with its major components are also identified by S. Rohatgi, [6]. Whereas, the implications of AND systems extend beyond these components. For instance:

(a) Their recommendation system often relies on content similarities, and metadata to suggest relevant resources to users. By incorporating AND techniques, the system can enhance its accuracy in recommending content authored by a particular individual.

(b) Similarly, the collaborator system facilitates scholarly collaboration by connecting researchers, authors, and institutions based on their academic contributions and interests. The AND system plays a crucial role in accurately identifying and linking authors to their respective publications and collaborations. This ensures that researchers can easily find and connect with potential collaborators with confidence, and accuracy.

(c) Similarly, researcher evaluation metrics, such as the H-index [7], which quantifies both the productivity and citation impact of the publications of a researcher, can be greatly improved with effective author name disambiguation. Which will foster effective research collaborations among institutions, funding agencies, and journals.

1.3 Motivation and Problem Formulation

Over the years a considerable[2] number of techniques have been proposed to perform AND, however, the precision of associating authors with their research papers is low. Though recall is not a major issue in AND, as, 100% recall can be achieved if all the name variants of the authors are available, however, the precision will be very low. In order to increase precision the recall usually gets compromised and

[2]Semanctic scholar search results against "author name disambiguation" resulted in more than 415 papers ranging from 2005-2022. This statistic is retrieved on 25^{th} January 2022

vice versa, which directly lowers the overall F1 score of the technique. For example, pairwise precision, recall, and F1 of a recent technique proposed by Pooja et al. [8], is 83.6%, 57.8%, and 62.1% respectively. A clear difference in the values of precision and recall scores can be seen. The reason of such low F1 score even in the presence of a reasonable precision can be contributed to its low recall. Moreover, the results of the technique proposed by Y. Zhang et al.,[9], gives a pairwise precision, recall, and F1 score of 77.9%, 63%, and 67.7% respectively, which also supports the fact that whenever the techniques try to raise the precision, the recall goes down.

Similarly, low precision even in the presence of high recalls often fails to give an overall reasonable F1 value. For example, a technique proposed by Liu et al [10] and Chen et al., [11] gives better recall as compared to precision i.e. (recall = 83% versus precision = 77%) and (recall = 90% versus precision = 86%) but a low F1 score of 79%, 88% respectively.

Therefore, there is a need to achieve better overall results without compromising the precision and with a balanced recall, as it is tried by Seol et al., [12] where the precision, recall and F1 scores are 94.7%, 94.8% and 94.7%. Similarly, by Peng et al., [13], where the pairwise precision, recall and F1 are equal to 78.6%, 71.8%, and 81.4%, respectively.

The second motivational issue is related to the automatic collection of data, which mostly use machine learning based techniques. These techniques need enriched datasets to establish rules, algorithms and training sets. For this purpose, there is a need to diagnose that the existing datasets of AND domain are sufficiently feature enriched, cover multidisciplinary scholarly data [14, 15], include data of authors who belong to different ethnic groups [16] such that the data is not skewed [15] with these aspects. There is a possibility that such datasets do not exist. Therefore, it is needed to explore this aspect as well.

Based on the discussed motivation, the following problem statement can be deduced:

"Existing author name disambiguation techniques have generally low precision, recall, or both, which affects their overall results. If a technique manages to raise precision, its recall often gets compromised, and if the recall is achieved its precision generally becomes low, which subsequently results in low F1 score."

1.4 Research Questions

Considering the problem statement, the following research questions have been devised:

1. How to devise an AND technique that can establish enhanced academic authorship's without compromising its precision? To devise such a technique: what feature combinations produce better results (in terms of F1 score) ?
2. How to curate an AND dataset which is feature enriched, covers multi disciplinary scholarly data and enclose authors belonging to multiple ethnic groups?

1.5 Objectives

Based on the research questions the following objectives have been formulated:

1. Formulate an AND technique that enhances the authorship results in terms of better F1 scores as compared to similar existing techniques, without compromising its precision.

   To achieve this: determine and identify feature ranking and combinations that can achieve better academic authorship's.

2. Use the identified feature ranking and feature combinations to formulate the AND technique.

3. Identify and collect a diverse range of scholarly data from various sources, ensuring feature en-richness, multi-disciplinary data coverage and multi ethnicity's of the authors. Validate the prepared dataset, and evaluate the proposed AND technique on the prepared dataset.

1.6 Scope of the Dissertation

Considering the objectives identified above, this section covers the scope of the study as follows:

1. To devise an AND technique that performs better in terms of F1 scores as compared to other similar techniques which include: [8, 17–22]
   (a) Compute feature rankings and identify feature combinations that give better F1 scores.
   (b) Devise an AND technique using the ranked features or better result producing feature combinations.
2. Dataset preparation
   (a) Identify gaps in the existing datasets ([14, 20, 23–27]) with respect to features availability, scholarly data disciplines and authors ethnic groups.
   (b) Prepare a dataset as per the identified gaps.
   (c) Evaluate the devised technique on the newly prepared dataset.

1.7 Research Methodology

A methodology is a system of principles, practices, and procedures applied to a specific branch of knowledge [28] to solve problems or add new knowledge. As shown in 1.4, a suitable methodology helps researchers to produce quality, valuable

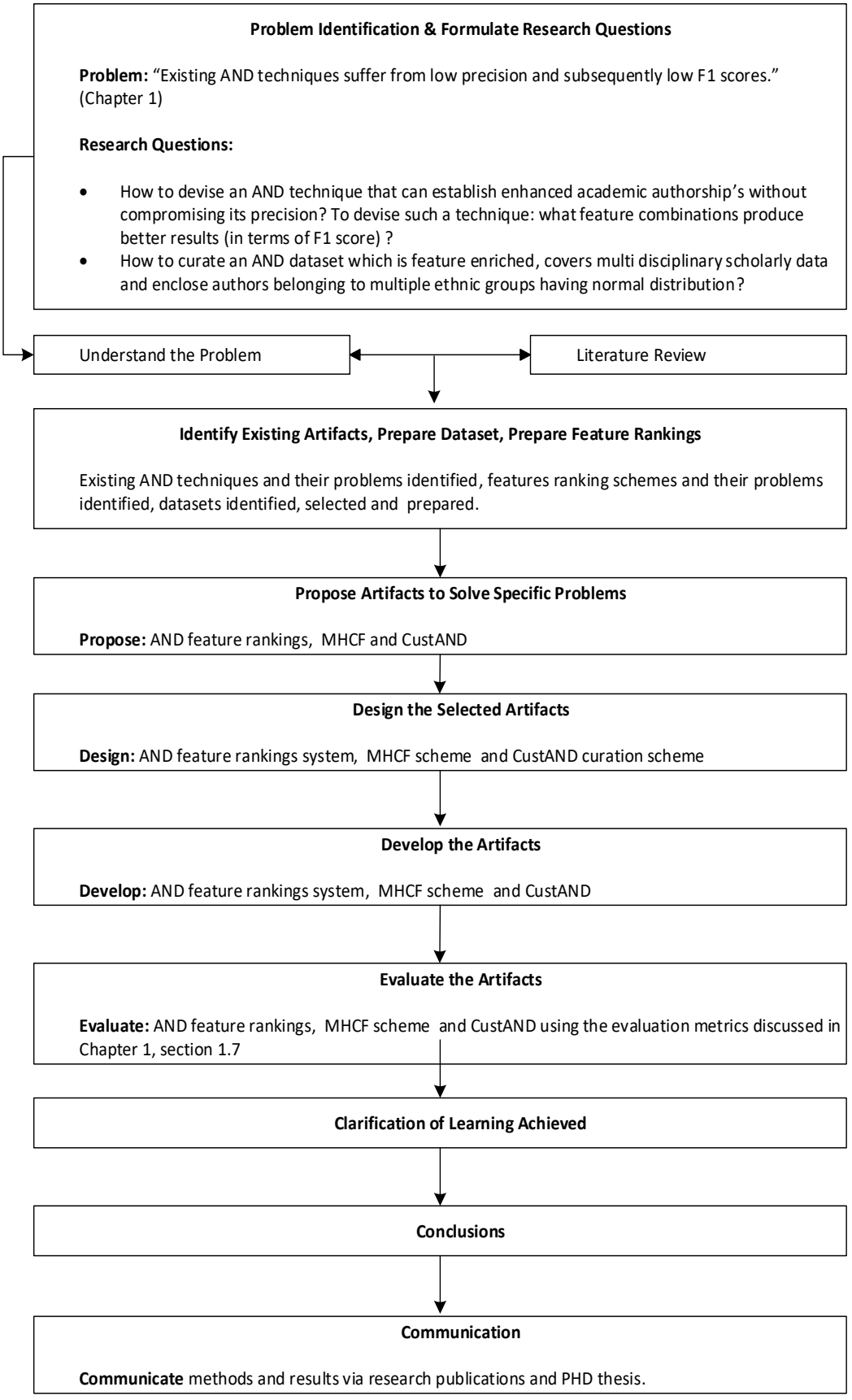

FIGURE 1.4: Research Methodology

and acceptable research that is publishable in research outlets. For this study, design science research methodology is adopted (DSR) [29] which is shown in Figure 1.4. In the first step, the problem is identified, research questions are formulated and objectives are defined. A deeper understanding of the problem is established by going through the existing literature. Existing techniques and their problems are identified, similarly, feature ranking techniques are identified and their problems are highlighted, also, different datasets from the literature are identified along with their problems, and datasets providing the most feature coverage are selected.

A new technique Multilayer heuristics based clustering framework (MHCF) is devised and developed to solve the problem with better results, which uses the feature rankings identified in this study. MHCF is developed and evaluated using standard metrics: pairwise precision, pairwise recall, and pairwise F1. The proposed technique is compared with several existing techniques with successful improvements. MHCF results are also reported using CustAND (which is curated and proposed in this study), with successful improvements as compared to the selected datasets (Arnetminer and BDBComp). Limitations of the research, improvements, and conclusions are reported in future work. Finally, the methods and findings are communicated to relevant audience through publications.

1.8 Evaluation Metrics

To measure the efficacy of the solution artifacts this study uses a variety of evaluation metrics that are commonly used in AND perspective. To measure the AND techniques results and ensure the validity, pairwise precision, pairwise recall and pairwise F1 scores are calculated. Similarly, average cluster purity, average author purity and K are also used, along with cluster precision, cluster recall and cluster F1 to measure the results, as identified by A.A. Ferreira et al [30]. For feature rankings, as identified by G. Chandrashekar et al and S. Alelyani et al[31, 32]. For

dataset's validity Cohen Kappa scores are used, as identified by J. Cohen et al [33]. Their details are given in chapter 4, section 4.3.3.

1.9 Dissertation Contributions

Considering the motivation and challenges outlined in section 1.2, the major research contributions of this study are presented as follows (also listed in table 1.1):

1. Multilayer Heuristics Based Clustering Framework (MHCF) for AND. (Chapter 4 elaborates the details).
   (a) Features ranking and better result producing feature combinations for author name disambiguation. (Chapter 3 discuss the details).
2. Completing features for author name disambiguation: an predicted analysis. (Chapter 5 elaborates the proposed dataset curation approach and its details).

TABLE 1.1: Contributions.

Research Questions (RQ)	**Contributions**
RQ1	MHCF[34]
RQ2	CustAND[15]

1.10 Dissertation Organization

The organization of the dissertation is as follows: chapter 2 critically analyses the literature to give supporting evidence related to the research questions. Chapter 3 discusses the contribution with respect to: the impact of features and their combinations on academic authorship results. The outcome of this chapter is feature rankings and feature combinations which give better F1 scores. Chapter

4 covers the details regarding the proposed approach to enhance author name disambiguation for academic authorships. Chapter 5 discusses the details of the proposed dataset for author name disambiguation. Whereas, the findings and future work are covered in Chapter 6.

Chapter 2

Literature Review

This chapter encloses a literature review of the existing author name disambiguation techniques for the past twenty years (2003-2023). To conduct the review, a methodological approach is followed to identify, select, review, and synthesize the research topic appropriately as suggested by Kitchenham et al [35]. The overall process which is adopted comprises of three phases as explained under, and, shown in Figure 2.1:

1. **Define Phase:** In this step, research questions (strings) need to be developed to define the methodology and guide the process.

   For this study, and, as the first phase, a single research question string "author name disambiguation" is considered to accommodate the maximum number of research papers.

2. **Conduct the Search:** The next phase involves conducting the search. For this purpose, the study performs two sub-tasks.

   (a) **Identification of the keywords:** For this purpose, keywords "author" AND "name" AND "disambiguation" are identified, which are a string of the general form, as suggested by Milano et al, [36].

(b) **Selection of the databases/digital libraries to conduct the search:** Four databases are chosen for this purpose which include, IEEE [1], ACM [2], Springer [3], and Elsevier [4]. Whereas, Google Scholar [5] is used for cross-referencing and inclusion of papers other than these databases. The main rationale for considering these databases is that they are widely used to search the relevant papers in other such studies and surveys. This is also referred by Raj et al and Elnabarawy et al, [37, 38], while following a similar approach for their study.

After performing these steps, different options are selected (depending on the database being used) to refine the search query. This needs to be done because of the varying nature of the databases' search filters. For IEEE, the search query is used on the "Metadata" option. For ACM, the query is run on the "title" option. Similarly, the query is run in "Computer Science" domain, where the paper's language is 'English'.

3. **Report the Results:** In this phase, the search query is run on different databases to find the potential papers, as shown in Figure 2.2. During the initial search, the number of papers that are retrieved varies, i.e. 26, 72, 48, 107 from ACM, IEEE, Springer, and Elsevier respectively. Next, the first inclusion criteria (IC 1): Year range (2003 - 2023), is applied, which did not affect the inclusion of the papers in the pool, as shown in Figure 2.2. This is because the author name disambiguation techniques range in these years only. Next, IC 2: Article type (conference, journal), is specified which filtered the papers as 14, 72, 45, and 95 for ACM, IEEE, Springer, and Elsevier. Next, IC 3: Computer Science and English, is defined, which lowered the paper count. After this, in the next stage, the paper titles and abstracts of the filtered papers are manually read to find relevant papers.

[1] IEEE digital library database provides access to documents related to the field of electrical, telecommunication, power and computer science literature.

[2] ACM (Association for Computing Machinery)

[3] Springer provides quality content related to science, technology and medical related fields, covering journals, conference proceedings, books etc.

[4] Elsevier ScienceDirect database give access to more than 2500 journals and 11000 books covering diverse fields.

[5] Google Scholar gives free access to full text or metadata of scholarly literature over a variety of disciplines and formats.

Among the filtered papers set, the papers are evaluated by topic and their importance to filter out highly important and partially important papers.

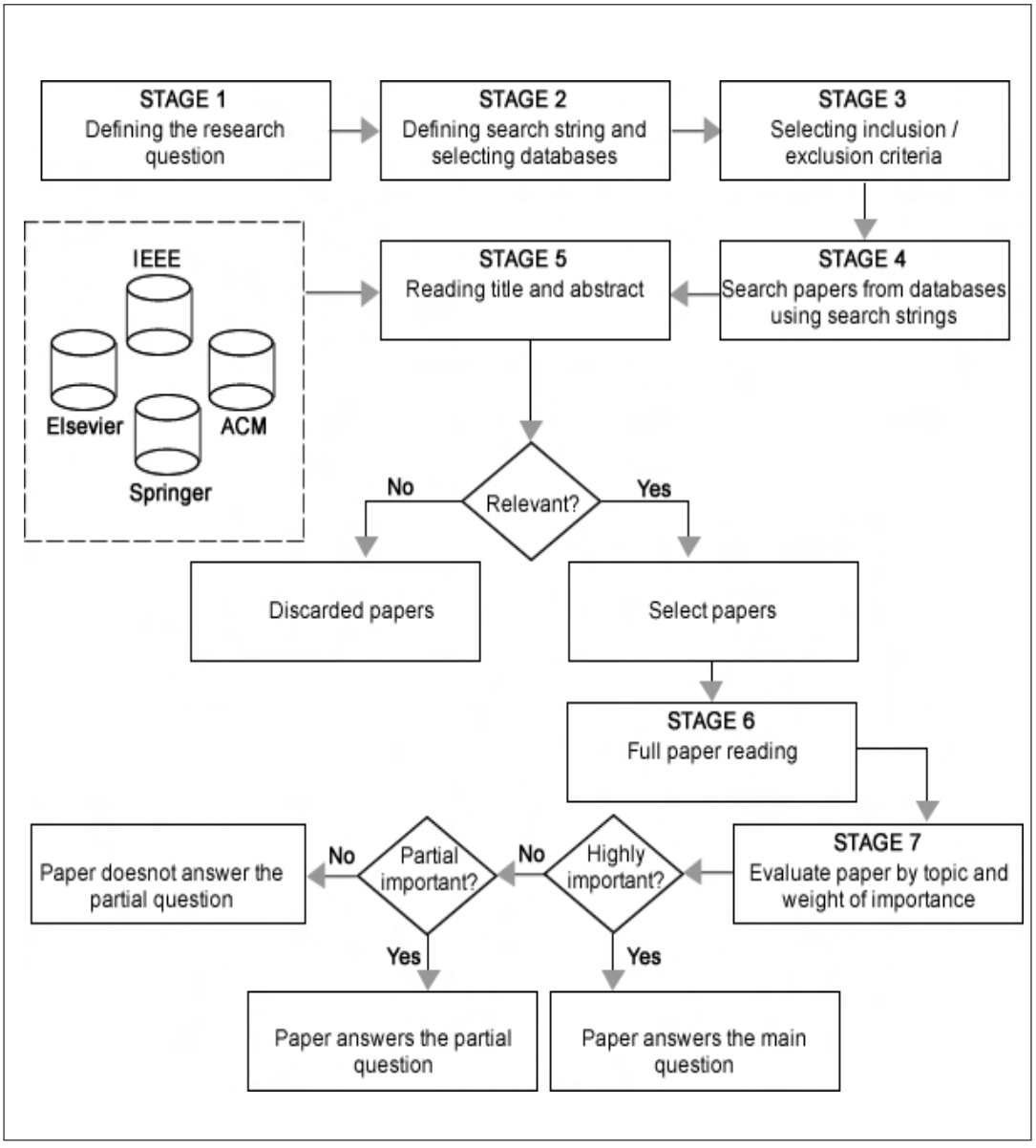

FIGURE 2.1: Literature Review Process

After completing these steps, Google Scholar (GS) is used to conduct the cross-referencing and selection of some other similar papers that belong to other journals, however, it is implied that cross-referencing and selection of papers from other databases is not part of the entire literature filtration process. The entire literature filtration process is shown in Figure 2.2.

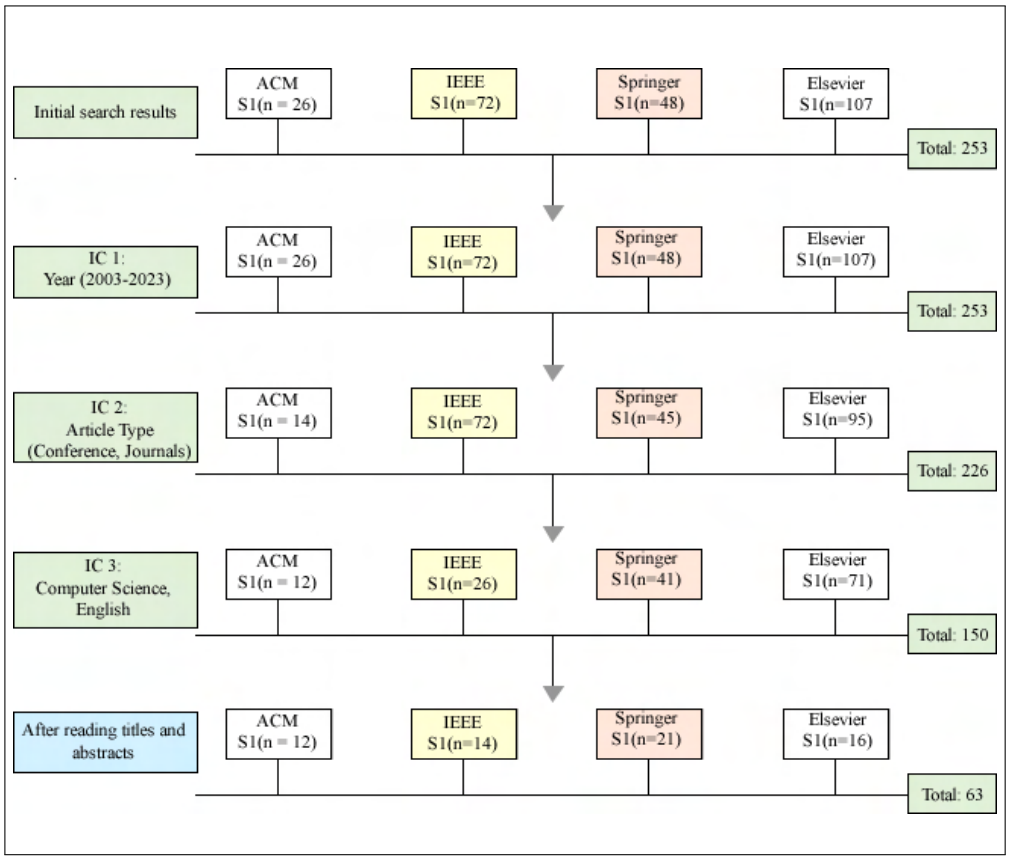

FIGURE 2.2: Literature Filtration Process

Figure 2.3 shows the graph of the filtered papers versus the selected paper's count following the above-explained criteria.

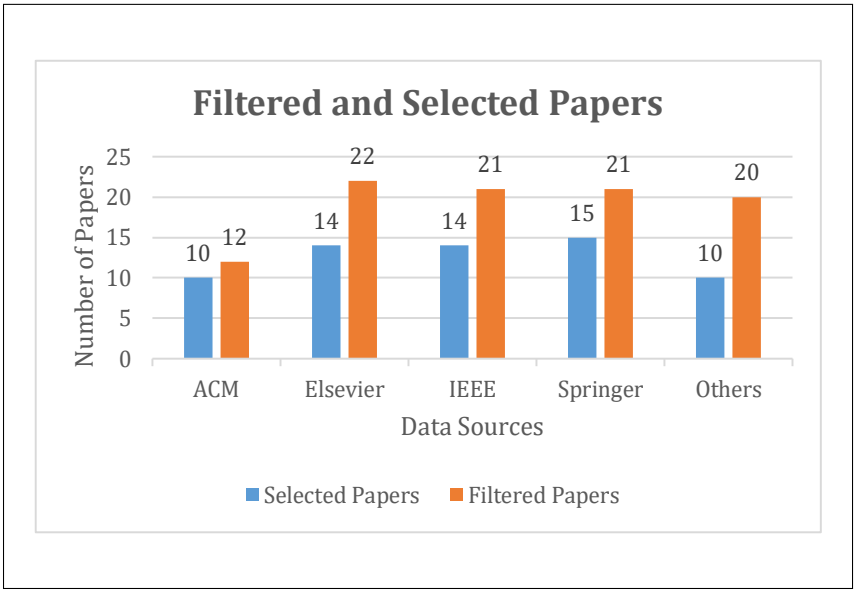

FIGURE 2.3: Statistics of Filtered Papers for Review

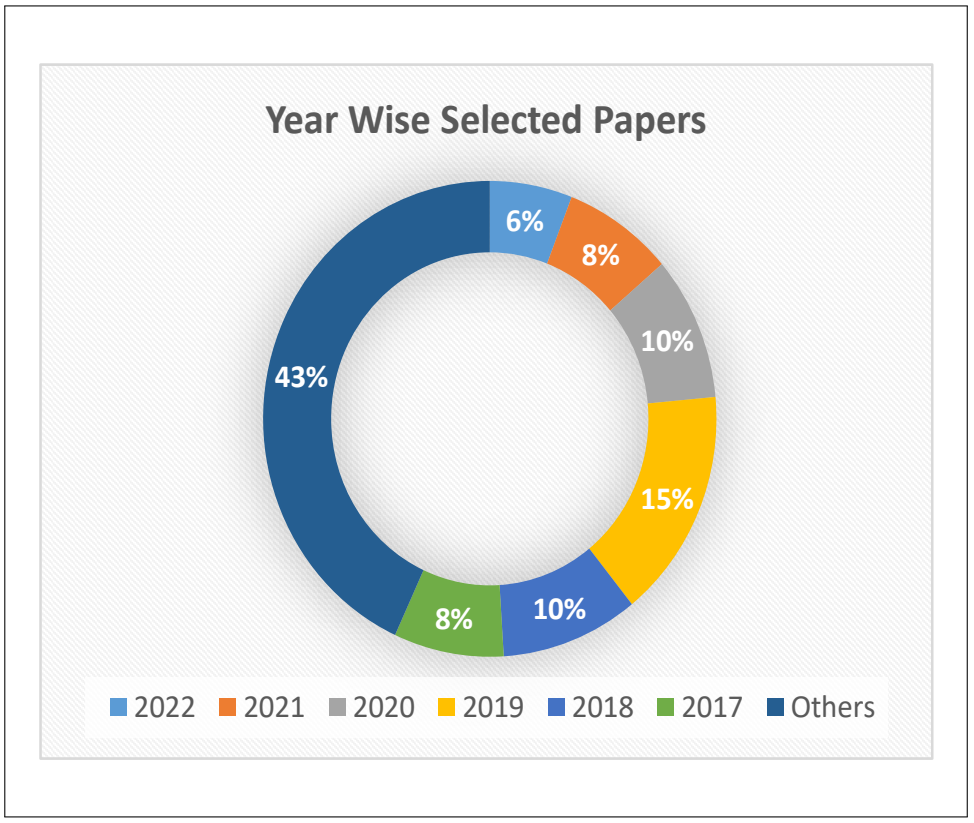

FIGURE 2.4: Selected Papers Distribution Year Wise

The year-wise distribution of the selected papers shows that more than 50% of the selected papers range in between the last 5 years, also shown in Figure 2.4.

Next, the selected papers are grouped under AND techniques related papers, features ranking related studies and datasets related papers. The techniques related papers are further grouped under supervised learning-based AND techniques, unsupervised learning-based AND techniques, graph-based learning techniques, as shown in Figure 2.5. The AND techniques are analyzed with respect to their methodology, the dataset, the features used by the technique, and their results.

The feature ranking papers are analyzed with respect to their feature ranking methodology, their proposed feature rankings and the dataset used by them.

For datasets related papers, the review examine the domain of the dataset, its labeling strategy, number of features available in them, and the ethnicity distribution of authors. In the end, this chapter converges the literature analysis and conclusion to the proposed problem statement.

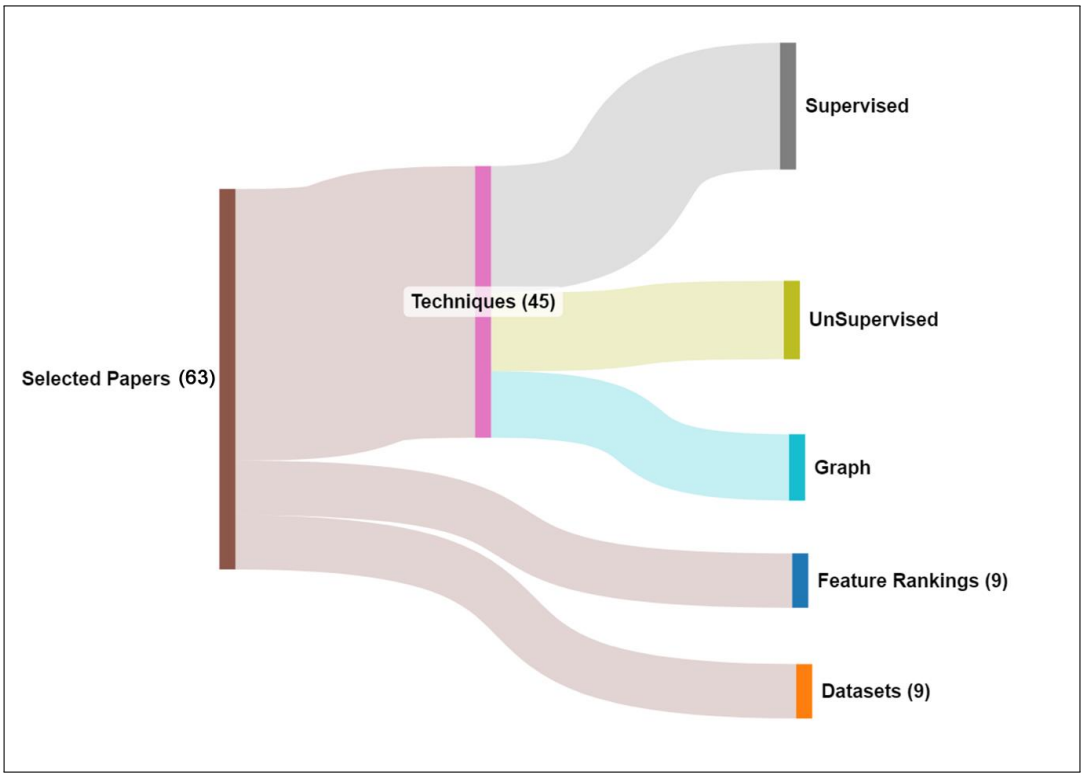

FIGURE 2.5: Selected Papers Grouping

2.1 Literature Review of Author Name Disambiguation Techniques

The following detail encloses a group-wise review of the existing techniques by considering their working, features, datasets, and their results. The literature review answers the following questions against each aspect which are outlined below:

1. **Working:**
   (a) Group the techniques as supervised, unsupervised, or graph-based learning method.
   (b) What is the working mode of the technique, i.e. batch or online?
2. **Features:**

   (a) What features are used by the reviewed technique?

   (b) Insight about the impact of features on the technique's results.

   (c) Insight about the usage of the useful[6] features by the existing AND techniques.

3. **Dataset(s):**

   (a) Which dataset(s) are used by the reviewed technique?

   (b) The labeling strategy adopted to label the dataset.

   (c) Insight of the feature's availability in the dataset(s).

   (d) Insight of the domain or the area of the publications in the dataset(s).

   (e) Insight of the authors ethnicity which is present in the dataset(s).

4. **Results:**

   (a) What are the results of the reviewed technique?

   The results are reported in the form of pP, pR, pF1, Precision (P), Recall (R), F1, ACP, AAP, K, accuracy, miss classification error, b^3 precision, b^3 recall, and b^3 F1.

The grouping of the techniques for this review is done as follows:

1. **Supervised Learning-Based AND Techniques (20 papers)**.

2. **Unsupervised Learning-Based AND Techniques (14 papers)**.

3. **Graph Learning-Based AND Techniques (11 papers)**.

4. **Studies that Analyzed the Impact of Features (9 papers):** These studies investigated the impact of different features on the performance of AND techniques. Table 2.4 comprehends them such that it holds the methodology adopted to evaluate the features, it includes a list of features that are declared to be useful to enhance the results, and the knowledge of

[6] features which contribute to make correct academic authorships.

whether the features are ranked explicitly or implicitly. The outcome of this particular analysis is a list of useful features which is listed in Table 2.5 such that the identified features are written in it without duplication.

2.1.1 Supervised Learning Based AND Techniques

Supervised learning based AND techniques use labeled data to train a model that predicts the correct author identity for a given paper. These techniques rely on existing knowledge to make accurate authorship determinations. The reviewed techniques are comprehended in the following Table 2.1.

TABLE 2.1: Overview of AND Techniques using Supervised Learning.

Ref	Features	Dataset	Results	Mode
[1]	Author forename, co-authors name, paper title, paper venue	DBLP	Precision= 70%, Recall= 95%, F1 = 90%	Batch
[11]	Author name, author affiliation, co-authors names, paper title, venue	Aminer, WhoIsWho	Precision=89.31%, 86.36%, Recall=80.80%, 90.33%, F1=84.84%, 88.3%	Online
[12]	Keywords, Email, Major of the author, Affiliation, co-authors names	Korean articles	Precision=94.78%, Recall=94.80%, F1=94.79%	Batch
[39]	Author name variants, co-authors names, paper title, venue	DBLP	Accuracy=73.3%	Batch
[40]	Author name variants, co-authors names, mesh word affiliations, publication year, paper title	Medline	Accuracy =95.99%	Batch

[41]	Co-authors, venues, keyword, paper title, abstract, area of interest	100 American authors	Miss classification error= 28%	Batch
[42]	Citing keywords, cited keywords, citing subject cat, addresses, cited subject cat, email, language, Cited journal titles, Author name initials	Web of Knowledge	b^3 F1=80.7%	Batch
[43]	Author name, affiliations, coauthor, author research interests, paper keyword	Vietnamese authors	Accuracy=99.31%	Batch
[44]	First author name, Paper title, Author affiliation, Publication venue, Co-authors list, Organization, Location, Email, Key phrases	PubMed	Precision=98.8%, Recall=96.3%, F1=97.5%	Batch
[45]	Author name variants, co-authors names, paper title, paper venue, keywords	Arnetminer	Mean F1 = 60.1%	Online
[46]	Author name, co-author name, paper title and venue) using BDBComp, (address, co-authors names, subject, reference paper, language, year, abstract and institution name) using WoS	BDB-Comp,WoS	Average Precision=92.7%, 88%, Average Recall=54.5%, 5%, Average F1 = 64.70%, 10%	Online
[47]	Author name variants, co-authors names, paper title, abstract, venue, publishing year	KISTI	Average MCC = 52.5%	Batch
[48]	Authors name variants, co-authors names, paper title, venue, user feedback	Arnetminer	F1=72.4%	Online

[49]	Author name variants, author affiliation, co-authors names, paper title, keywords, venue	Aminer, Pubmed	Average pairwise Recall = 69.1%, 89.2%, Average pairwise Precision=65.2%, 86%, Average F1=67%, 89.2%	Batch
[50]	Author first and last names, name initials, affiliation, organization, publication year, email, location, co-authors, journal and semantic types	Medline	Precision=82.7%, Recall=92.2%, F1=87.2%	Batch
[51]	Author name, co-authors names, paper title, venue	Aminer	Average Precision= 87.93%, Average Recall= 77.74%, Average F1= 82.53%	Online
[52]	Author name, author affiliation, paper title	Aminer	Precision=76.92%, Recall= 64.54%, F1 = 70.19%	Batch
[53]	Author name, co-author names, affiliation, paper title, abstract, venue, year	Aminer, WhoIsWho	Macro pairwise Precision=78.22%, 79.96%, pairwise Recall=56.04%, 50.50%, pairwise F1=65.3%, 61.90%	Batch
[54]	Author full names, co-author names, paper title, venue	DBLP	Precision=98.9%, Recall=99.1%, F1=98.8%	Batch
[55]	Author name variants, co-author names, paper title, year	Dutch data repository	F1=75.43%	Batch

[56]	Paper title, Sources title, author first name, co-authors, venue	CiteSeerX, DBLP (5 names)	Macro F1=71.3%, micro avg F1 = 51%	Batch

Though a lot of supervised learning-based techniques exist that try to learn a classification model for authors, where it is considered that they can be highly precise when they are trained with adequate, high-quality labeled data. However, all of this requires tedious human efforts. Moreover, likely, these methods often fall prey to imbalanced class problems as active researchers generally have more publications and relatively more training data as compared to the non-active ones. The results of these techniques also point out this fact, as, majority of the reviewed techniques have low precision, recall, and overall F1 scores. It is evident from the results that such techniques usually cause false class predictions when deployed in real environments due to the daily increase in researchers and their research publications.

2.1.2 Unsupervised Learning Based AND Techniques

Unsupervised learning based techniques do not use labeled data, but instead, learn to identify authors by clustering papers based on their features. Existing AND studies which use unsupervised learning to establish academic authorship's are comprehended in Table 2.2.

TABLE 2.2: Overview of AND Techniques using Unsupervised Learning.

Ref	**Features**	**Dataset**	**Results**	**Mode**
[10]	Co-authors name, paper title, paper venue	DBLP	Average pairwise Precision= 77.9%, pairwise Recall= 83.9%, pairwise F1 = 79%	Batch

[13]	Author affiliation, co-authors names, paper title, summary, venue, publication year	Arnetminer, DBLP, CiteSeerX	(Diting) Precision=78.6%, 82.2%, 66.4%, Recall= 71.8%, 85.4%, 60.1%, F1 =74.5%, 83.2%, 63.5% (Diting++) Precision=85.3%, 84.6%, 74.4% Recall=73.8%,89.6%, 68.4%, F1=81.4%, 87.1%,71.2%	Batch
[18]	Email, Co authorships, Paper title	11 ambiguous authors names	Pairwise F1=75%	Batch
[20]	Author name variants, affiliation, co-authors names, venue	DBLP, BDBComp	Pairwise Precision=84%, 67%, Pairwise Recall= 65%, 79%, Pairwise F1 = 73%, 71%	Batch
[57]	Author name variants, co-authors names, paper title, venue	DBLP	Accuracy ranges between 61.5% to 64.7%	Batch

[58]	Author full name, author surname, author initials, publication page, abstract, title and co-authors	BT digital library, Web	Accuracy =73.33%	Batch
[59]	Author name variants, affiliation, co-authors names, venue	BDBComp, Synthetic dataset	Average Cluster Purity (ACP)=99.7%, 82.1%, Average Author Purity (AAP)=77.2%, 71.9%, K=87.7%, 76.8%	Online
[60]	Author name variants, co-author names, paper title, venue	KISTI, BDBComp	Pairwise Precision=93.5%,86.8%, Pairwise Recall=97.4%,,84.1%, pairwise F1 = 95.4%, 85.5%	Online
[9]	Paper title, abstract, co-authors names, venue, affiliations	Aminer	Precision= 77.96%, Recall=63.03%, F1=67.79%	Batch
[61]	Co-authors, paper title, venue, abstracts, author affiliation	Arnetminer, DBLP, CiteSeerX	Macro F1=74.5%, 83.2%, 63.5%	Batch
[62]	Author name variants, author affiliation, co-authors names, paper title, keywords, venue	Chinese authors	Precision=95%, Recall= 96%, F1 = 95%	Batch

[63]	Author name variants, author email, co-authors names, self citation	WoS	Pairwise F1=90%	Batch
[64]	Author name variants, author affiliation, co-authors names, paper title, abstract, venue	Aminer, CiteCeerX, DBLP	Macro pairwise F1=84.7%, 68%, 88%	Batch

Unsupervised learning-based techniques try to place papers written by distinct authors into their respective clusters, such that each cluster belongs to one author. These techniques work without labeled training data, relying on patterns and structures within the data itself. They are beneficial when labeled data is scarce or when it's impractical to manually label large datasets. e.g. cases like inactive or new researchers versus active or old (researchers who have been actively involved in research for quite some time). However, the requirement of prior knowledge of the number of authors or k partitions is a major drawback faced by such techniques. Though some researchers catered this problem by following heuristics-based clustering approach but most of them use counting the most frequent words using binary, Term Frequency Inverse Document Frequency (TF IDF), Cosine, Jaccard similarity measurement strategies to analyze and cluster publications, which is ineffective and often raise cluster impurities. The effects of the adoption of such strategies are evident from their AND results, with low precision, recall, and F1 scores as shown above.

2.1.3 Graph Based Learning AND Techniques

These techniques use graph-based methods to represent the relationships between authors and papers and then use these representations to disambiguate authors. The reviewed techniques are comprehended in Table 2.3.

TABLE 2.3: Overview of AND Techniques using Graph based Learning.

Ref	Features	Dataset	Results	Mode
[8]	Co-authors name, paper title, abstract, venue, affiliation, reference	Arnetminer, Aminer	Pairwise Precision=83.6%, 60.9%, Pairwise Recall=57.8%, 59.9%, Pairwise F1=62.1%, 55.4%	Batch
[17]	Co-authors names, paper title, venue	DBLP, Arnetminer	Pairwise F1 = 71% , 82%	Batch
[61]	Co-authors names, affiliation, paper title, paper summary, venue	Arnetminer, DBLP, CiteSeerX	Macro F1= 74.5%, 83.2%, 63.5%	Batch
[65]	Co-authors names	DBLP, PubMed	Average Precision=94.1%, 100%, Average Recall=83%, 96.4%, Average F1=86.1%, 98%	Batch
[66]	Author name variants, co-authors names	Arnetminer,CiteSeerX	Average pairwise F1 = 81.6%, 63.8%	Batch
[67]	Co-authors names, paper title	Arnetminer-S, Arnetminer-L	Pairwise F1 = 84%, 80%	Batch

[68]	Co-authors name, paper title, venue, year	CiteseerX	Macro F1= 62.1%	Batch
[69]	Author name, co-authors name, affiliation, paper title, keywords, abstract	Aminer	F1 = 60.2%	Batch
[70]	Author name, co-authors, affiliation	Aminer	Average Precision=78.1%, Average Recall=67.47%, Average F1=72.40%	Batch
[71]	Co-authors name, meta-content (paper title, abstract, venue, reference etc)	Arnetminer, Aminer	Pairwise Precision=72.4%, 75.6%, Pairwise Recall=75.1%, 67.1%, Pairwise F1=71.5%, 69.4%	Batch
[72]	Co-authors name, meta-content (paper title, abstract, venue, reference, year)	Arnetminer, DBLP	Pairwise Precision= 73.8%, 82.2%, Pairwise Recall=67.2%, 69.1%, Pairwise F1=68.4%, 71.5%	Online

Graph-based learning techniques utilize different relationships in the data to form academic networks, and based on these relationships try to perform author name disambiguation. These techniques represent authors as nodes and their relationships, like co-authorships and citations, as edges. This approach proves beneficial in capturing the intricacies of academic connections, even when dealing with limited data points. However, since these techniques rely on specific features to make heterogeneous networks, this leads to the incapability of resolving certain common use cases. For instance: most of graph-based techniques use co-authors' names to make a co-author network, however, this leads to the incapability of resolving single-authored publications, and, publications with more co-authors, as the co-author names can also be ambiguous. Thus, increasing false positives in the results. This observation is visible through these technique's low precision, recall, and F1 scores as given in the table above. Some researcher's besides working on enhancing the author name disambiguation process in academic authorship's, also diverted their attention to study and rank the features that are used in this process. This is an important aspect to be considered besides enhancing the working of the technique, as relevant and independent inputs can contribute positively to increase the true positives in the results [15, 50]. Table 2.4 reviews such studies, whereas Figure 2.6 shows the Sankey[7] visualization of the reviewed studies.

TABLE 2.4: Overview of Studies that Evaluate the Impact of Features on the Authorship Results.

Ref	**Features evaluated**	**Methodology**	**Useful features**	**Ranking**
[1]	Full forename, co-authors, title, venue	String based matching, algorithmic disambiguation methods	Full forenames	Implicit
[12]	Keywords, Email, Major of the author, Affiliation, Common co-authors	SVM based classifier is used to evaluate different feature combinations.	Co-authors, Email, Keywords, Major of the author, Affiliation	Implicit

[7] A visualization to view the flow of information in a process or system also discussed by [73]

[18]	Email, Co authorships, Paper title	Package-merge clustering is used to evaluate different features.	Email, Co authorships, Paper title	Implicit
[40]	Author first name, middle name, last name (idf), Author suffix, Author order, Affiliation (softtfidf, tfidf, Jaccard), Co-author last name (shared, idf, jacquard), Mesh shared (idf, tree, tree shared idf), Journal shared (idf), Journal language (idf), Journal year, Journal year (difference), Title shared	Feature ranking is based on RFs permutations and Gini factor along with correlations between features and their output class.	1. Author last name (idf), 2. Author middle name, 3. Affiliation (tfidf), 4. Journal year, 5. Affiliation (softtfidf), 6. Mesh shared (idf)	Explicit
[42]	This study used 1,080 features in total (18 features from literature, 3 citation metadata, 24 citing and cited features and 1,035 conjunctive features)	Authors evaluated combination of features to evaluate how well they perform at bootstrapping phase using high precision rules.	1. Citing keywords, 2. Cited keywords, 3. Citing subject cat, 4. Addresses, 5. Cited subject cat, 6. Email, 7. Language, 8. Cited journal titles, 9. Author name initials	Explicitly mentioned top speed optimized features

[44]	First author name, Paper title, Author affiliation, Publication venue, Co-authors list, Organization, Location, Email, Key phrases	Different combinations of subsets of important features are evaluated using RF, C4.5, KNN and SVM. The study tried to assess the impact of feature sets on the scheme's performance.	First author name, Paper title, Publication venue, Co-authors list, Organization, Location, Email, Key phrases	Implicit
[50]	Author first and last names, initials, affiliation, type of organization (university, hospital, research center etc.), publication year, email, location, co-authors, journal descriptors and semantic types.	Decision tree algorithm	1. Journal descriptors, 2. Semantic types, 3. co-authors, 4. Ambiguity score, 5. First name, 6. Last name length, 7. Years difference, 8. City, 9. Type of organization, 10. Language, 11. Country, 12. Initials, 13. Affiliations, 14. Email	Explicit
[57]	Author first name initial and last name. Author first name with first three characters, Author full names, Co-author name, Publication title, Publication venue, Author research areas	Used K-way spectral clustering. The study compared single features with each other.	Author full names, Author first name with first three characters, Co-author name	Implicit

[58]	Author full name, Author surname, Author initials, Co-authors, Paper title, Paper abstract, Author publication web page	The study used proton ontology's-based instance unification approach. Different feature combinations are assessed following different similarity measures.	Paper title, Paper abstract, Author full name, Author initials	Implicit

After considering the review of these studies (Table 2.4), it is observed that to the best of the knowledge, few studies explicitly ranked set of features, whereas, some

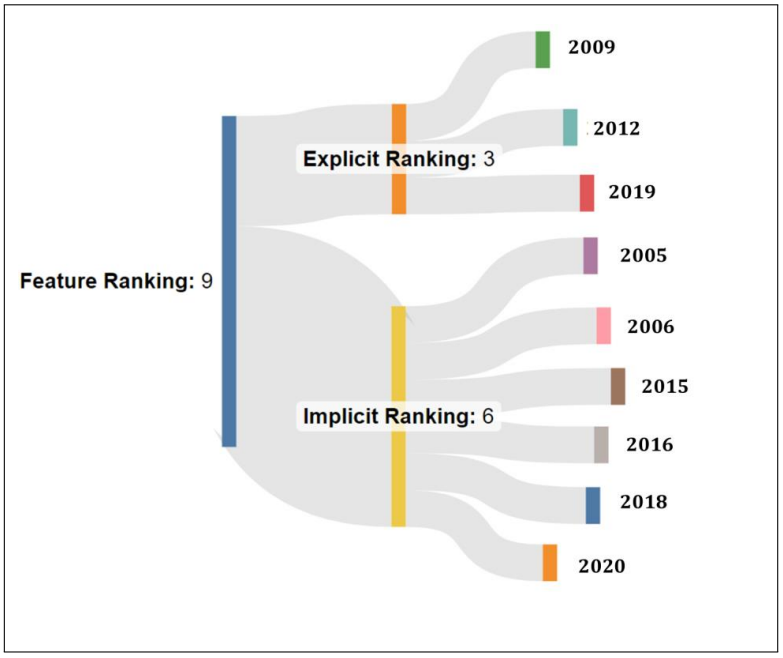

FIGURE 2.6: Sankey view of Feature Ranking Studies with Publishing Year

studies implicitly talked about the features based on their experimental results. This review in general identifies a set of useful features that can contribute positively to make correct authorships. However, they lack a complete picture in this regard. A list of all the useful features that are identified by the nine reviewed studies is given in Table 2.5 without duplication. Whereas, the complete analysis of this aspect is covered in section 2.2.2.

After the identification of useful features, a complete review of their presence in the existing publicly available datasets is carried out. The datasets included in this review are those which are commonly used by the existing AND techniques and are publicly available. The review of the datasets is covered in two tables, where Table 2.6 enlists the details covering the area or domain of the data included in the dataset, the labeling strategy (LStrategy) adopted to label the data in it and the percentage of the ethnicities of the authors present in them. Whereas, the availability of the 17 useful features in the reviewed datasets is covered in Table 2.7. Similarly, the Sankey visualization of them is given in Figure 2.7, which highlights the domain of the dataset, its labeling strategy, and the number of features in them.

TABLE 2.5: Useful Features Declared in Literature.

No	**Feature Name**	**No**	**Feature Name**
1	Author name variants	10	Journal language
2	Author Email	11	Year of publication / publication time
3	Author Affiliation	12	Addresses
4	Author Research Interests / field / subject categories	13	Cited Journal titles
5	Co-author names	14	Cited article keywords
6	Title of paper	15	Citing article keywords
7	Abstract of paper	16	Cited publication subject category
8	Keywords of paper	17	Citing publication subject category
9	Publication venue		

TABLE 2.6: Review of AND Datasets.

No	Dataset	Area	LStrategy	Ref	Ethnicity of Authors
1	DBLP	CS	Manual	[39, 57]	Indian (10%), Chinese (36%), English (15%), Japanese (3%), Korean (38%)
2	BDBComp	CS	Manual	[20]	Hispanic (100%)
3	Arnetminer	CS	Manual	[25]	Indian (9%), Chinese (62%), English (22%), German (4%), Hispanic (3%)
4	KISTI-AD-E-01	CS	Manual	[26]	Indian (9%), Chinese (47%), English (12%), German (5%), Hispanic (5%), Arab (1%), Dutch (1%), French (1%), Greek (1%), Israeli (2%), Italian(2%), Japanese (2%), Korean (11%), Nordic (1%), Salv (1%)
5	PubMed	Medical	Manual	[44]	Indian (20%), Chinese (9%), English (40%), German (9%), Hispanic (8%), Italian (5%), African (3%), Japanese (3%) and Arabs(3%)
6	Aminer	CS	Manual	[9]	Chinese (95%), English (1%), Hispanic (1%), Korean (1%),
7	Medline	Medical	Manual	[50]	East Asian origin (85%) (i.e. China, Japan, Mongolia, North Korea, South Korea, Taiwan)

8	Pubmed	Medical	Semi-Automatic	[74]	Indian (20%), Chinese (9%), English (40%), German (9%), Hispanic (8%), Italian (5%), African (3%), Japanese (3%) and Arabs(3%)
9	Medline	Medical	Semi-Automatic	[74]	East Asian origin (85%) (i.e. China, Japan, Mongolia, North Korea, South Korea, Taiwan)

TABLE 2.7: Useful Features Coverage by AND Datasets (datasets are mentioned by their numbers as represented in Table 2.6 due to space limitation).

	Features	**Datasets**									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Total**
1	Author Full Name	✓	✓	✓	✓	X	✓	X	✓	X	6
2	Author First Name	X	X	X	X	X	X	✓	✓	✓	3
3	Author Middle Name	X	X	X	X	X	X	X	✓	✓	2
4	Author Last Name	X	X	X	X	X	X	X	✓	✓	2
5	Author Short Name	X	✓	X	✓	✓	X	X	X	✓	4
6	Author Email	X	X	X	X	X	X	X	✓	✓	2
7	Author Affiliation	X	X	✓	X	X	✓	X	✓	✓	4
8	Author Research Interests	X	X	X	X	X	X	X	X	X	0
9	Co-author names	✓	✓	✓	✓	X	✓	X	✓	✓	7
10	Title of paper	✓	✓	✓	✓	X	✓	X	✓	✓	7
11	Abstract of paper	X	X	X	X	X	✓	X	✓	✓	3

12	Keywords of paper	X	X	X	X	X	X	X	X	X	0
13	Publication venue	✓	✓	✓	X	X	✓	X	✓	✓	6
14	Journal language	X	X	X	X	X	X	X	X	X	0
15	Year of publication	X	X	✓	✓	X	✓	X	✓	✓	5
16	Addresses	X	X	X	X	X	X	X	X	X	0
17	Cited Journal titles	X	X	X	X	X	X	X	X	X	0
18	Cited article keywords	X	X	X	X	X	X	X	X	X	0
19	Citing article keywords	X	X	X	X	X	X	X	X	X	0
20	Cited publication subject cat	X	X	X	X	X	X	X	X	X	0
21	Citing publication subject cat	X	X	X	X	X	X	X	X	X	0
	Total	**4**	**5**	**6**	**5**	**1**	**7**	**1**	**11**	**11**	**-**

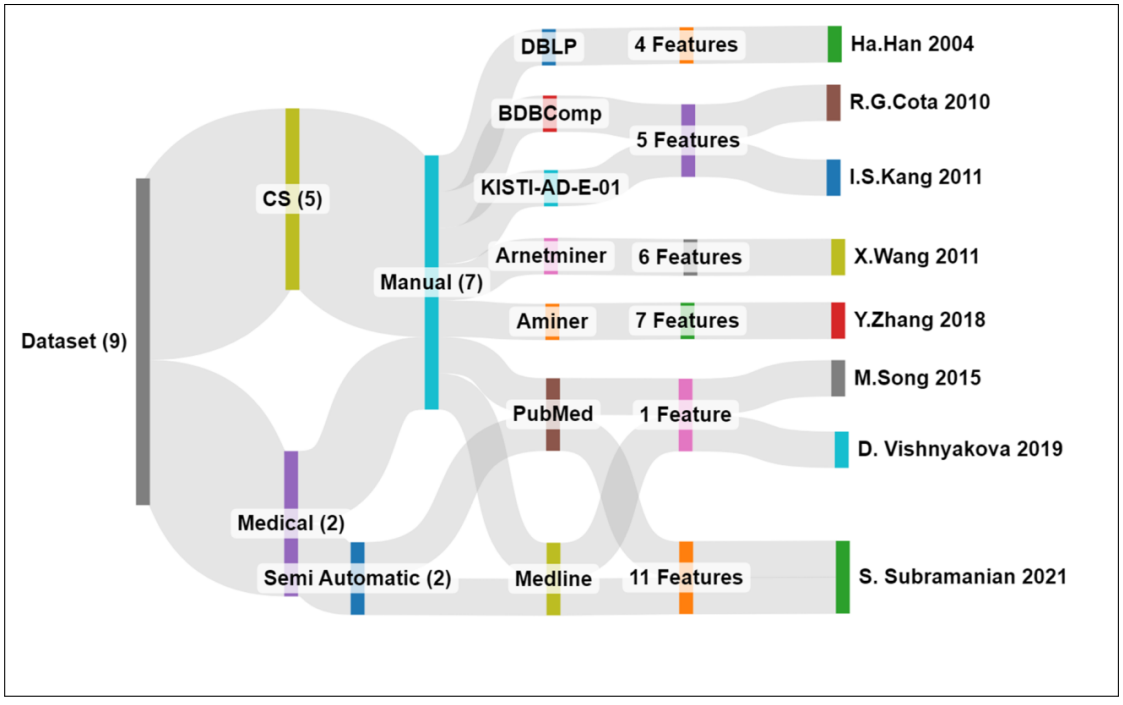

FIGURE 2.7: Reviewed Datasets

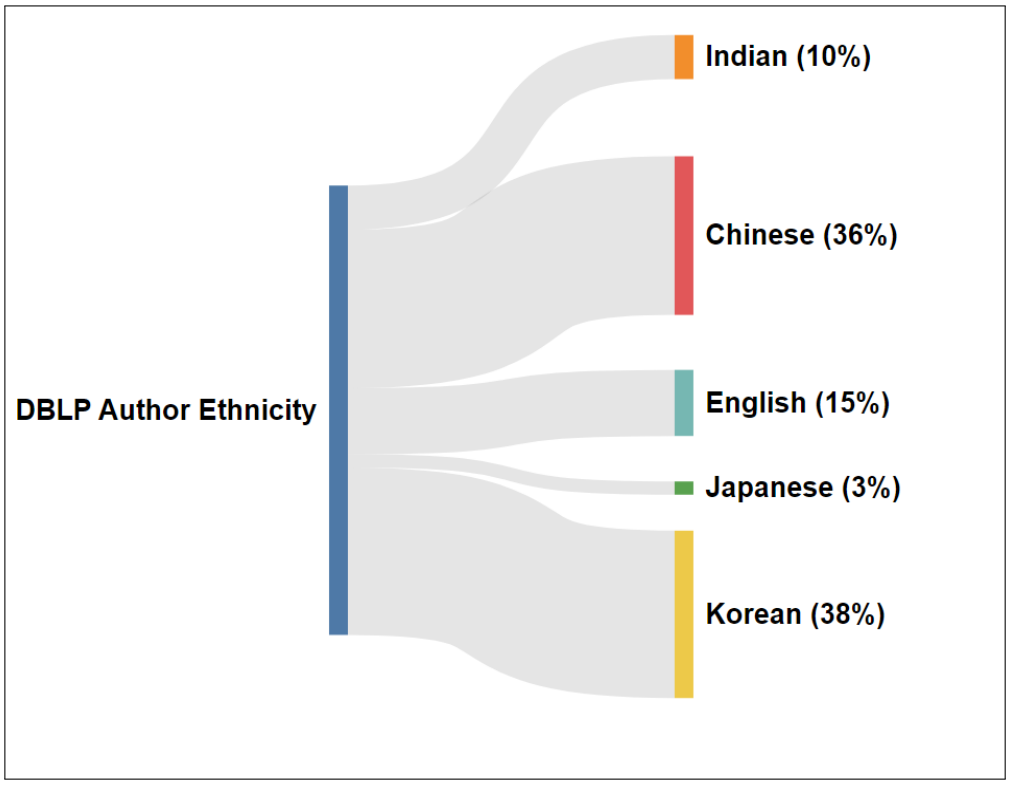

FIGURE 2.8: Author Ethnicity in DBLP Dataset [57, 75]

FIGURE 2.9: Author Ethnicity in BDBComp Dataset [20]

The author ethnicity distribution in DBLP and BDBComp datasets is shown in Figure 2.8 and 2.9 respectively.

Figure 2.10 and 2.11 shows the Sankey diagrams of the author ethnical groups in Arnetminer and KISTI-AD-E-01 datasets respectively. Similarly, Figures 2.12,

2.13, and 2.14 show the author ethnicity distributions in PubMed, Aminer and Medline datasets respectively.

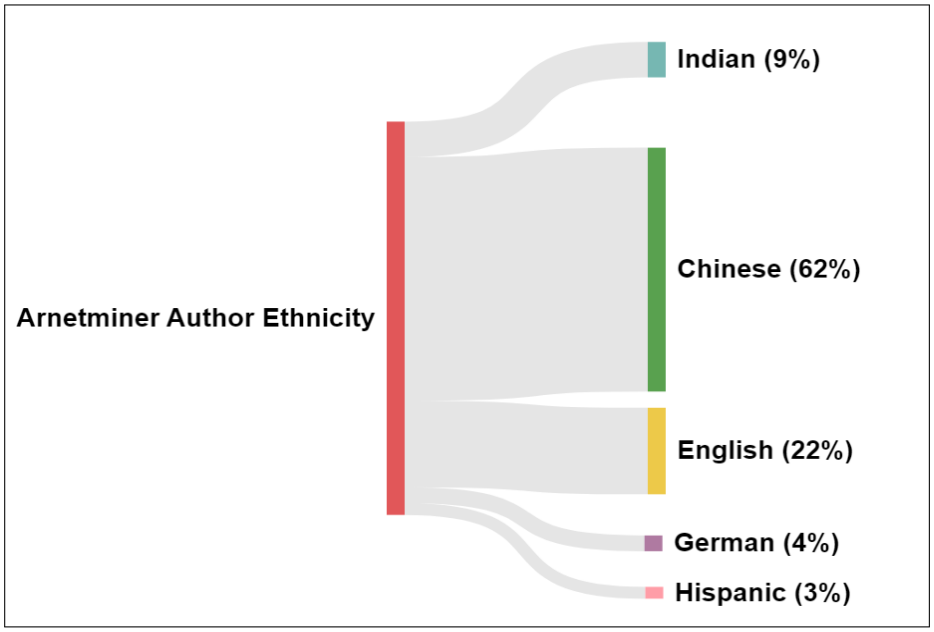

FIGURE 2.10: Author Ethnicity in Arnetminer Dataset [25]

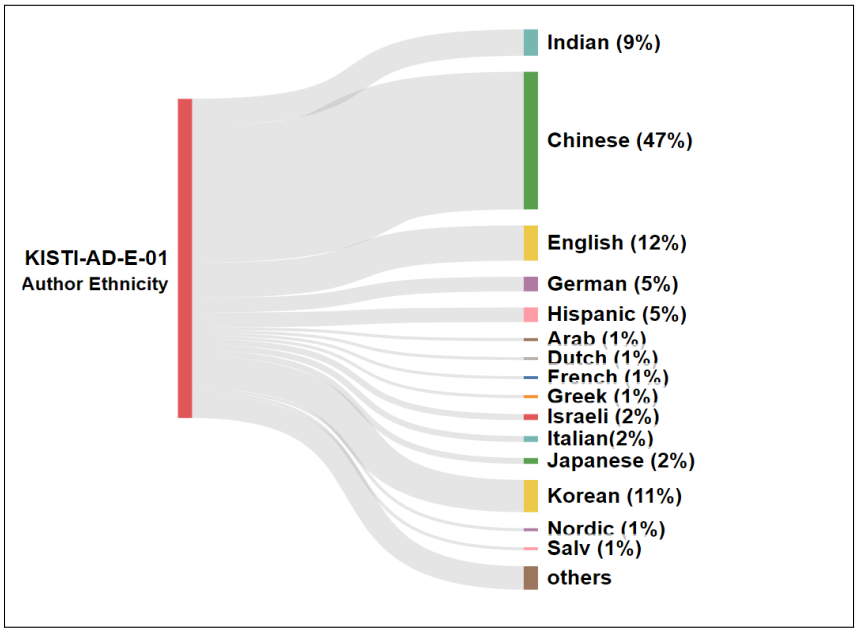

FIGURE 2.11: Author Ethnicity in KISTI-AD- E-01 Dataset [26]

A comprehensive review of the existing AND techniques is also carried out which focuses on observing the usage of the 17 useful features in them. The review also

takes into account the results reported by these techniques, which can partially identify the relationship between the features and their impact on the results. Table 2.8 gives a comprehensive overview in this regard.

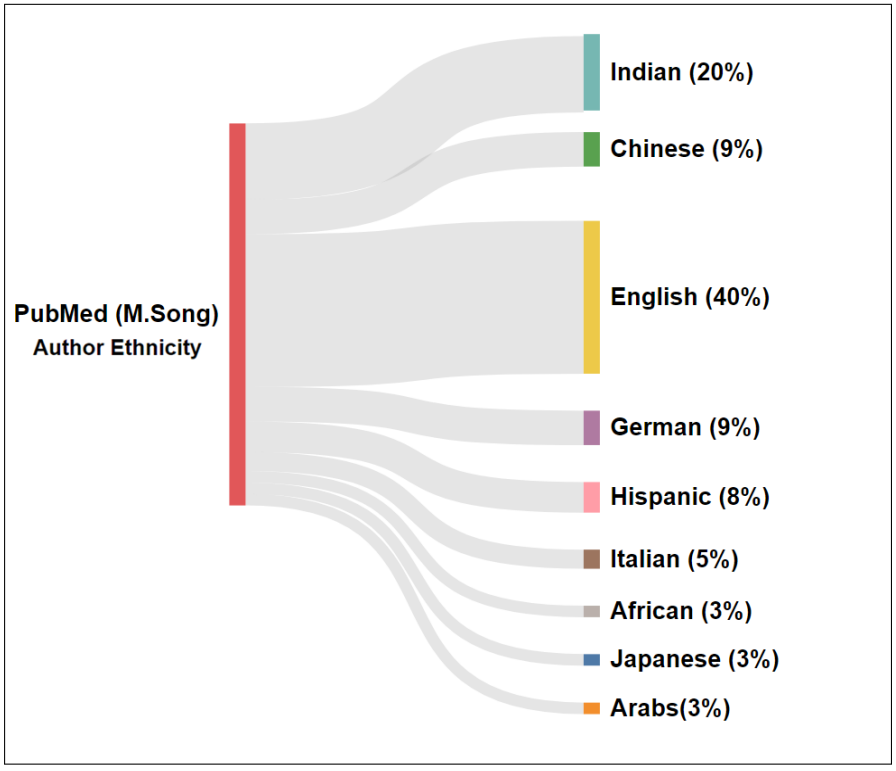

FIGURE 2.12: Author Ethnicity in PubMed Dataset [44, 74]

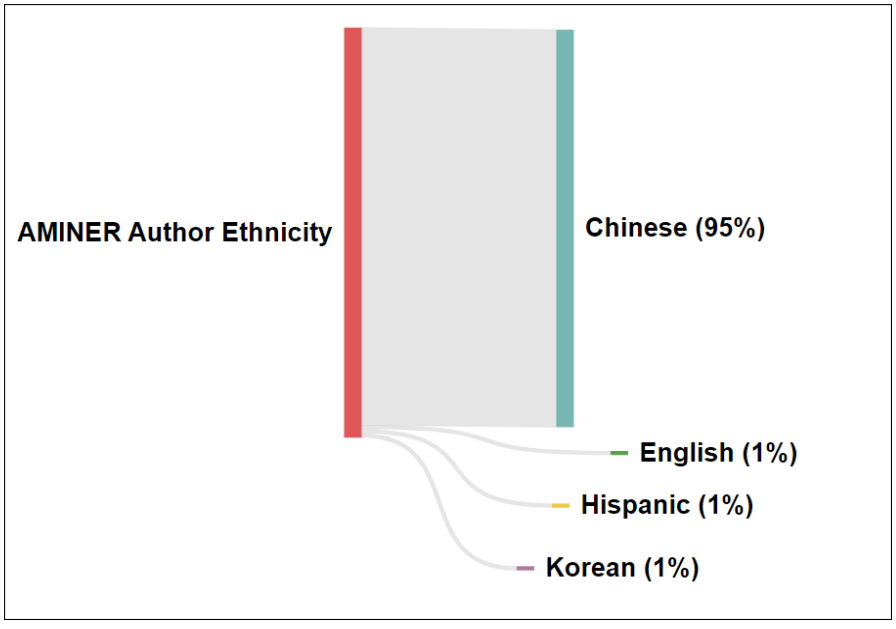

FIGURE 2.13: Author Ethnicity in Aminer Dataset [9]

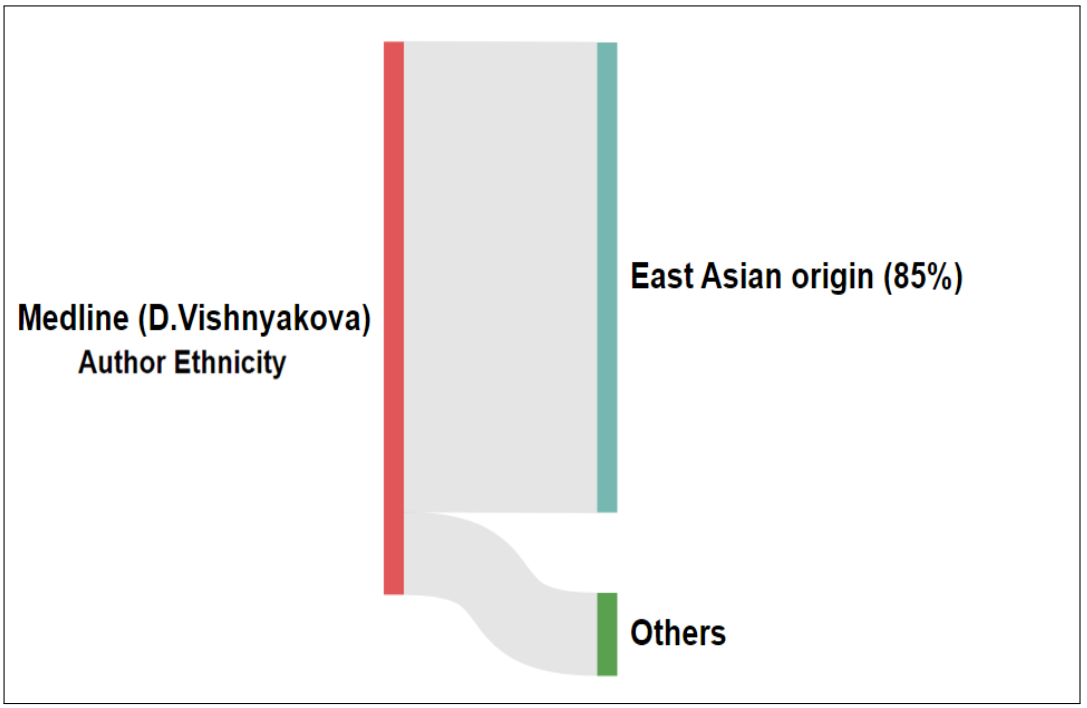

FIGURE 2.14: Author Ethnicity in Medline Dataset [50, 74]

The next section covers a detailed analysis of the literature review. The analysis is comprised of two subsections. The first section analyzes the reviewed techniques with respect to their results i.e. precision, recall, and F1 scores. This will complement in developing an understanding of the problems in the existing techniques and provide evidence regarding the research question 1 which is raised in Chapter 1. The analysis extends further to identify the problems in the existing techniques which focus on the features ranking, and better results producing feature combinations aspect. This section also highlights the problems and potential gaps in the existing reviewed datasets, focusing on the availability of features in them, the domain of the data that is present in the dataset, and the ethnicity of the authors whose data is present in them, in particular.

The analysis of these aspects is necessary to later select appropriate existing datasets to address the features ranking and feature combinations related gaps, along with their use to develop and test the proposed author name disambiguation technique.

TABLE 2.8: Overview of AND Techniques with Respect to Feature Usage.

	Features																	**Results**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	16	**17**	
[1]	✓	X	X	X	✓	✓	X	X	✓	X	X	X	X	X	X	X	X	F1 = 60% - 90%
[8]	✓	X	✓	X	✓	✓	✓	X	✓	X	X	X	✓	✓	X	X	X	pF1= 62.1%, 55.4%
[10]	X	X	X	X	✓	✓	X	X	✓	X	X	X	X	X	X	X	X	pF1 = 79%
[12]	X	✓	✓	✓	✓	X	X	✓	X	X	X	X	X	X	X	X	X	F1=94.79%
[13]	✓	X	✓	X	✓	✓	✓	X	✓	X	✓	X	X	X	X	X	X	F1 = 81.4%, 87.1%, 71.2%
[17]	X	X	X	X	✓	✓	X	X	✓	X	X	X	X	X	X	X	X	pF1 = 71% , 82%
[20]	✓	X	✓	X	✓	X	X	X	✓	X	X	X	X	X	X	X	X	pF1 = 73%, 71%
[22]	✓	X	X	X	✓	✓	X	X	✓	X	X	X	X	X	X	X	X	pF1 = 79.6%, 75.2%
[25]	✓	X	✓	X	✓	✓	X	X	✓	X	X	X	X	X	X	X	X	Avg F1 = 89.20%
[39]	✓	X	X	X	✓	✓	X	X	✓	X	X	X	X	X	X	X	X	Acc = 73.30%
[40]	✓	X	✓	X	✓	✓	X	X	X	X	✓	X	X	X	X	X	X	Acc= 95.99%
[42]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	$B^3F1 = 80.7\%$
[44]	✓	✓	✓	X	✓	✓	X	✓	✓	X	X	X	X	X	X	X	X	F1=97.5%

[46]	✓	✓	X	X	✓	✓	X	X	✓	X	X	X	✓	X	X	X	X	Avg F1 = 64.70%, 10%
[47]	✓	X	X	X	✓	✓	✓	X	✓	X	X	X	X	X	X	X	X	Avg MCC = 52.5%
[48]	✓	X	X	X	✓	✓	X	X	✓	X	X	X	X	X	X	X	X	F1 = 72.4%
[49]	✓	X	✓	X	✓	✓	X	✓	✓	X	X	X	X	X	X	X	X	Avg F1 = 67%, 89.2%
[50]	✓	✓	✓	X	✓	X	X	X	X	X	✓	X	X	X	X	X	X	F1 = 87.2%
[52]	✓	X	✓	X	X	✓	✓	✓	✓	X	X	X	X	X	X	X	X	F1 = 70.19%
[53]	✓	X	✓	X	✓	✓	✓	X	✓	X	✓	X	X	X	X	X	X	pF1=65.3%, 61.90%
[54]	✓	X	X	X	✓	✓	X	X	✓	X	X	X	X	X	X	X	X	F1=98.8%
[57]	✓	X	X	X	✓	✓	X	X	✓	X	X	X	X	X	X	X	X	Acc = 61.5% - 64.7%
[58]	✓	X	X	X	✓	✓	✓	X	X	X	X	X	X	X	X	X	X	Acc = 73.30%
[59]	✓	X	✓	X	✓	X	X	X	✓	X	X	X	X	X	X	X	X	K= 87.7%, 76.8%
[60]	✓	X	X	X	✓	✓	X	X	✓	X	X	X	X	X	X	X	X	pF1 = 95.4%, 85.5%
[9]	X	X	✓	X	✓	✓	✓	X	✓	X	X	X	X	X	X	X	X	F1 = 67.79%
[61]	X	X	✓	X	✓	✓	✓	X	✓	X	X	X	X	X	X	X	X	macro F1=74.5%, 83.2%, 63.5%
[64]	✓	X	✓	X	✓	✓	✓	X	✓	X	X	X	X	X	X	X	X	F1=84.7%, 68%, 88%
[66]	✓	X	X	X	✓	✓	X	✓	✓	X	X	X	X	X	X	X	X	mean F1=60.1%

[68]	X	X	X	X	✓	✓	X	X	✓	X	✓	X	X	X	X	X	X	F1 = 62.1%
[70]	✓	X	✓	X	✓	X	X	X	X	X	X	X	X	X	X	X	X	F1 = 72.40%
[71]	X	X	X	X	✓	✓	✓	X	✓	X	X	X	X	X	X	X	X	pF1=71.5%, 69.4%
[72]	X	X	X	X	✓	✓	✓	X	✓	X	X	X	X	X	X	X	X	pF1=68.4%, 71.5%
[75]	✓	X	X	X	✓	✓	X	X	✓	X	X	X	X	X	X	X	X	Acc = 85.35%
[76]	X	X	X	X	✓	X	X	X	X	X	X	X	X	X	X	X	X	Avg F1= 86.1%, 98%
[77]	✓	X	✓	X	✓	X	X	X	✓	X	X	X	X	X	X	X	X	K=88.8%, 72.7%, 87%
[78]	✓	X	✓	✓	✓	X	X	✓	X	X	X	X	X	X	X	X	X	Acc = 99.31%
[79]	✓	X	✓	✓	✓	X	X	X	X	X	X	X	X	X	X	X	X	F1 = 77.9%
[80]	✓	X	X	X	X	✓	✓	X	✓	X	✓	X	X	X	X	X	X	F1 = 85.66%
[81]	✓	✓	✓	X	X	X	X	X	X	X	✓	X	X	X	X	X	X	F1 = 85.56%
[82]	✓	X	✓	X	✓	X	X	X	✓	X	X	X	X	X	X	X	X	F1 = 98.87%, 96.78%
[83]	✓	✓	X	X	✓	X	X	X	X	X	X	X	X	X	X	X	X	pF1=90%
[84]	✓	X	✓	X	✓	✓	X	✓	✓	X	X	X	X	X	X	X	X	F1 = 95%
[85]	✓	X	X	X	✓	✓	X	X	✓	X	X	X	X	X	X	X	X	F1=98.8%
	36	8	23	4	43	33	13	8	34	1	8	1	3	2	1	1	1	-

2.2 Analysis of the AND Techniques

This section includes a critical analysis of the AND techniques which are reviewed in the previous section.

2.2.1 Analysis with Respect to the Results

This section includes group-wise distribution graphs to visualize the results (F1 scores) of the reviewed author name disambiguation techniques (i.e. supervised, unsupervised, and graph-based techniques). In the upcoming analysis, only the highest F1 score-achieving techniques are discussed per group.

Supervised learning based techniques:

The result distribution of supervised learning based techniques is shown in Figure 2.15, which shows that only three techniques' F1 scores lie in the range of 91.1% - 98.8%. One of the techniques with high F1 scores is proposed by J.W. Seol et al [12], where, its F1 score is 94.79%, the precision score is 94.7% and the recall score is 94.8%. The results of the technique are reported on a handful of Korean authors, which makes the dataset author ethnic-centric [15, 16], and is not available for further evaluation. Moreover, the technique uses binary similarity measures to identify distinct authors using the paper title, co-authors, and keyword comparisons. However, such comparisons can fail to capture situations in which keywords are not exactly the same, there are no common co-authors, publications are single-authored or the co-authors' names are not full [34]. The high precision and recall scores of this technique might be compromised if the technique is tested on a dataset that is not author-centric and has more generalized scenarios included in it.

Another AND technique with high F1 score is proposed by M. Song et al [44], which gives a high F1 score of 97.50%. The dataset used in it is PubMed, whereas the features include author's first name, paper title, publication venue, co-authors, organization, location, email entities, and keywords. PubMed is a domain-centric

dataset, as it only deals with the medical-related papers. This means the technique is designed to cater author name ambiguity problem with this domain perspective and might suffer from false positives when applied to a generalized dataset. Also, the single-authored publication scenarios will be hard to cater by the technique, as the dataset used to design the technique has the least number of such cases [50].

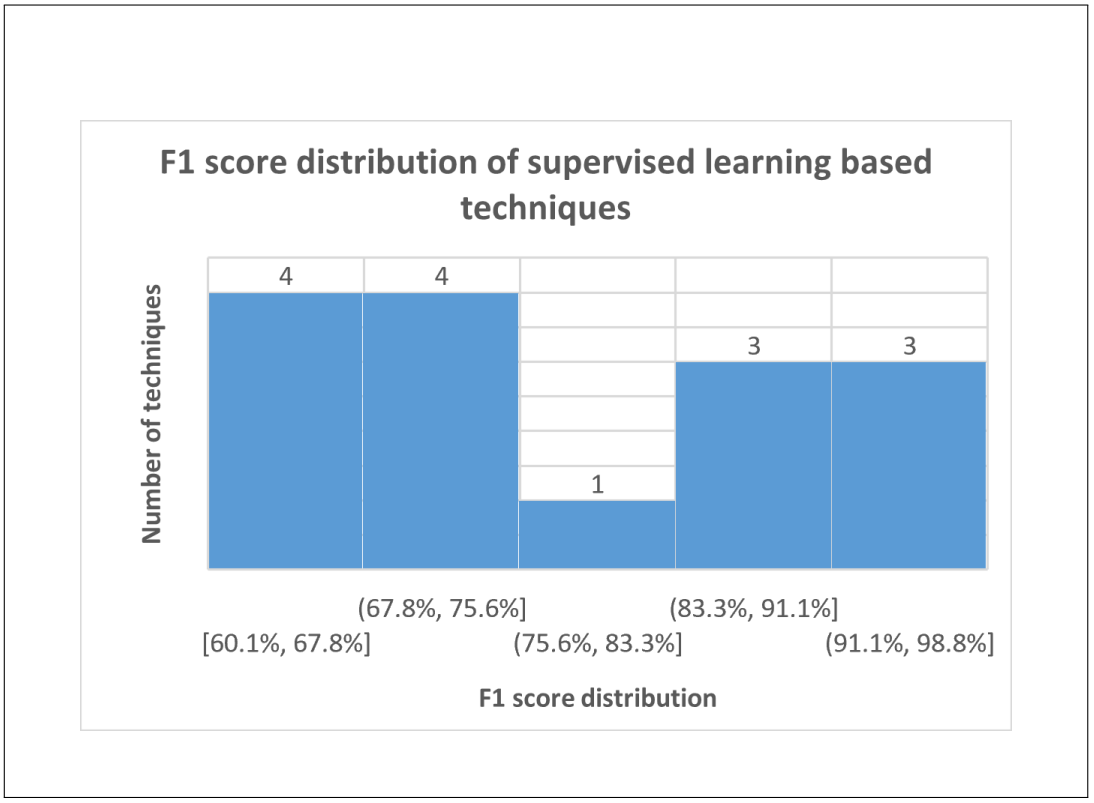

FIGURE 2.15: F1 Score Distribution of Supervised Learning-Based Techniques.

The author name disambiguation technique proposed by Z. Boukhers et al, [54] is Bib2Auth, which is based on a deep learning based model, with a high F1 score of 98.80%, precision score of 98.9% and a recall score of 99.1%. The model uses different attributes like the author's full name, co-author's name, paper title, and paper venue to perform academic authorships. One reason for such a high F1 score can be the use of the author's full name as a feature which is a good feature unless it is shared among distinct authors. Another reason is that it only considers a small part of the DBLP[8][9] repository, to test their technique's results, whose accuracy is

[8]It is an online reference of bibliographic information on computer science publications.
[9]https://dblp.org/faq/How+accurate+is+the+data+in+dblp.html

itself not guaranteed, as mentioned by the maintainers of the platform. Moreover, DBLP[10] also holds domain-specific data i.e. Computer Science domain. Also, the technique will face difficulty in case it encounters variation in paper title words when an author co-authors with a completely different author or works in cross domains.

Unsupervised learning-based techniques:

The F1 score distributions of the unsupervised learning-based techniques are shown in Figure 2.16.

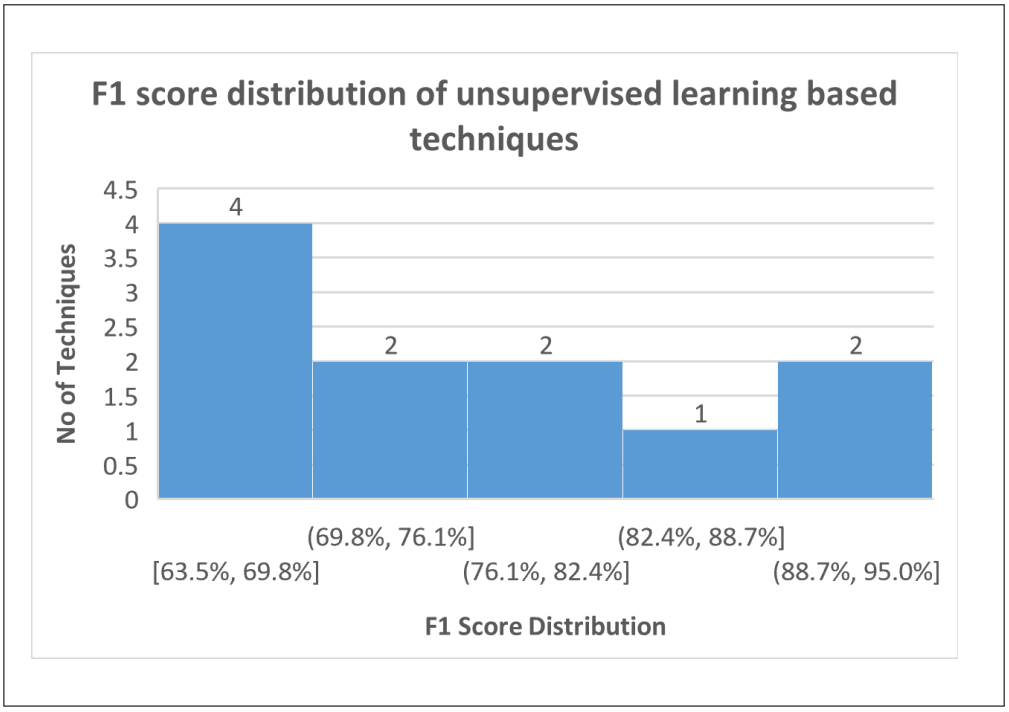

FIGURE 2.16: F1 Score Distribution of Unsupervised Learning-Based Techniques.

The distribution clearly points out that only two techniques in this group have high academic authorship results i.e. their F1 scores lie in the range of 88.7% - 95%. The techniques are proposed by J. Kim et al [63], with an average pairwise F1 score equal to 90%, and S. Zhang et al [62] AND technique with a high F1

[10]It is an online reference of bibliographic information on computer science publications.

score of 95%. J. Kim et al [63] proposed technique, uses a dataset which is based on Web of Science (WoS) articles published between 2012 and 2016 in different top 100 Computer Science (CS) journals. However, the dataset details are not given, whereas it is also noted that it is CS area specific, which lacks many common patterns occurring in other areas [16, 34]. Because of this, and the high probability of encountering ambiguous co-authors' names in other domains, it is more likely that this technique will encounter high false positives, in the real world. Similarly, though the AND technique proposed by S. Zhang et al [62] gives high precision, recall, and F1 scores to encounter author name ambiguity problems, i.e., 95%, 96%, and 95% respectively, the reports are based on Chinese names, which makes the dataset ethnic-centric. Because of this, the technique might not be able to give such high results if tested using a generalized dataset with this respect or when applied to real-world data.

Graph based techniques:

The graph based techniques result distribution can be visualized in Figure 2.17. It is evident that majority of the reviewed techniques F1 score distribution range between 60.2 - 65.4% which is low. Similarly, only five techniques F1 score range between 70.6% - 85.1%, which has quite some room for improvement. The graph based technique proposed by X. Fan et al [65] has an average F1 score = 86.1% which is reported on DBLP using only co-authors name. This score can be improved further by utilizing other features as well. Though the same technique is reported to attain a high average F1 score of 98% using PubMed dataset that too using a single feature (co-authors names), however it is worth noting that the papers published in Medical domain have greater number of co-authors as compared to other domains. Because of this the high F1 scores are obtained. However, this particular methodology might not produce such high results when applied on any other domain other than medicine. Figures 2.18 (better precision vs compromised recall and its effect on F1 score) and 2.19 (better recall vs precision and its effect on F1 score) specifically show this observation. In the above-discussed techniques,

the precision, recall, and F1 scores are observed to lie in approximately the same range. However, most of the other techniques encounter a common phenomenon, i.e. if their precision increases, the recall gets compromised, whereas, if the recall improves the precision gets compromised, thus impacting the overall F1 scores.

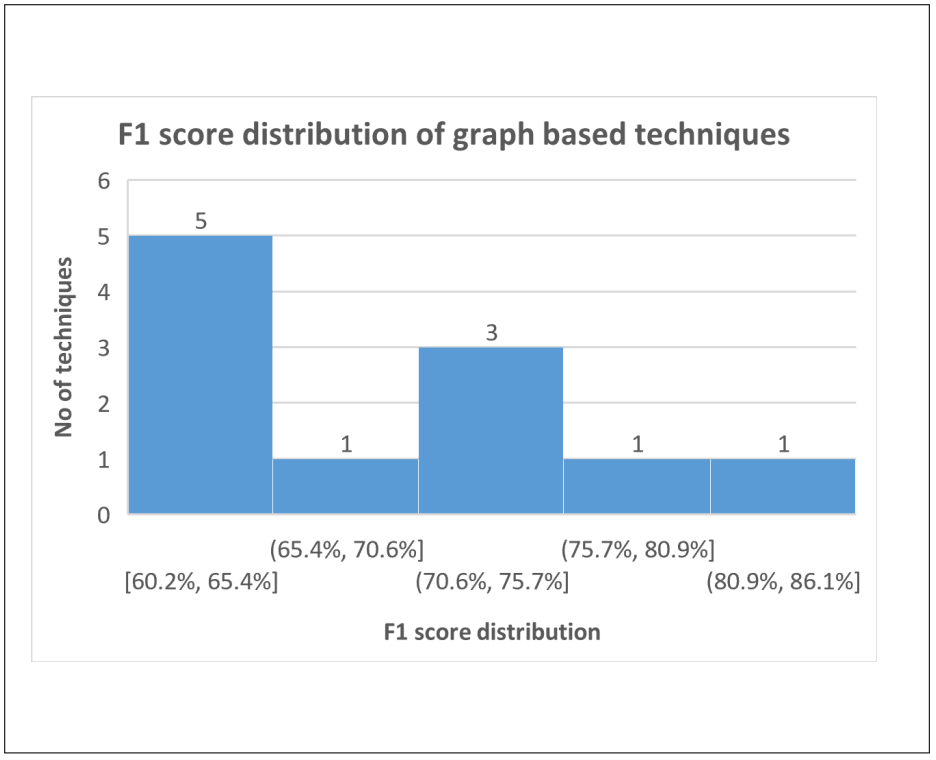

FIGURE 2.17: F1 Score Distribution of Graph-Based Techniques.

For example in Figure 2.18 the technique proposed by Z. Zhang et al. [52] gives a slightly better precision score of 76.92%, as compared to its recall of 64.54% but the overall F1 score is low i.e. 70.19%. Similarly, Y. Chen et al., [53], give a slightly better precision score of 78.22% in comparison to the recall scores of 56.04%, but the overall F1 score is quite low i.e. 65.3%. Likewise, Q. Sun et al., [51] gives a better average precision of 87.93% versus an average recall of 77.74%, B. Chen et al., [11] give a precision, recall, and F1 of 89.31%, 80.80%, and 84.84%, using Aminer dataset. Pooja et al proposed one batch and one online technique i.e., [8, 72], where the precision score of their batch-based technique is better i.e. 83.6% as compared to their recall i.e., 57.8% with a low F1 score of 62.1%. Whereas, their online technique's precision, recall, and F1 scores are almost in the same range, i.e., 73.8%, 67.2%, and 68.4%. In the literature, there are some other

techniques in which the recall scores are slightly better than the precision scores however, better recall in the absence of reasonable precision also fails to give an overall reasonable F1 value.

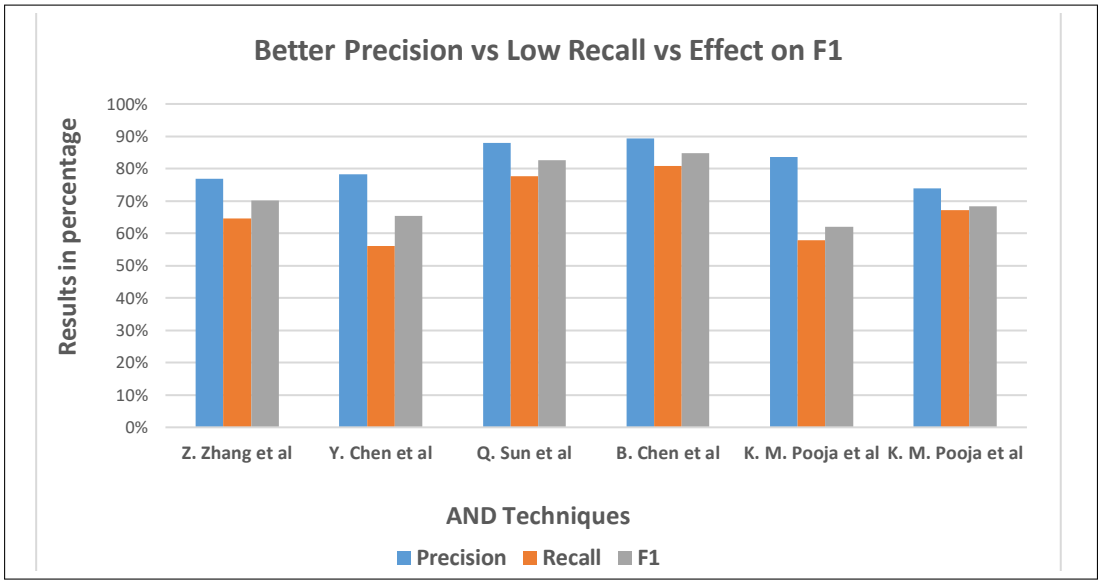

FIGURE 2.18: AND Techniques with Better Precision versus Low Recall.

This behavior can be visualized in Figure 2.19. For example, a technique proposed by Liu et al., [10] gives a better recall score of 83% as compared to the precision score of 77% and an F1 score of 79%. Similarly, B. Chen et al., [11] proposed technique's recall score is 90% versus a precision score of 86% and an F1 score of 88%. Also, K.M. Pooja et al proposed technique's recall score equal to 75.1%, a precision score equal to 72.1%, and F1 equal to 71.5%. From this analysis, it can be seen that if a technique manages to raise precision, its recall often gets compromised, and if the recall is achieved its precision generally becomes low, which subsequently results in low F1 scores.

To summarize the analysis, a conclusive graph is presented in Figure 2.20, illustrating the F1 score distribution of all reviewed author name disambiguation (AND) techniques. The distribution of these scores indicates that, in general, most AND techniques exhibit low precision, recall, and F1 scores, highlighting their limited ability to accurately disambiguate academic authorships.

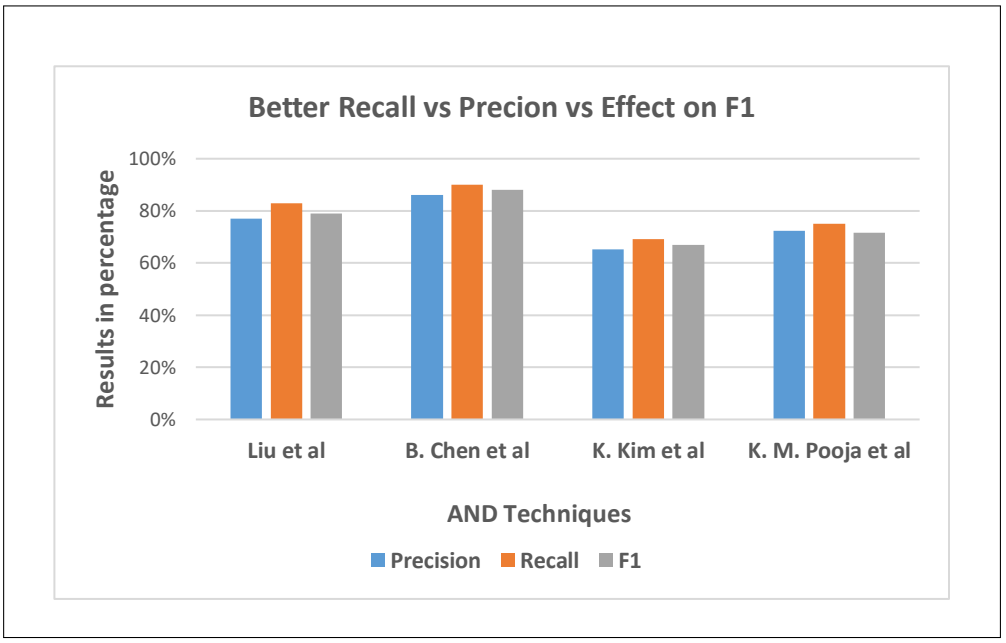

FIGURE 2.19: AND Techniques with Better Recall versus Precision.

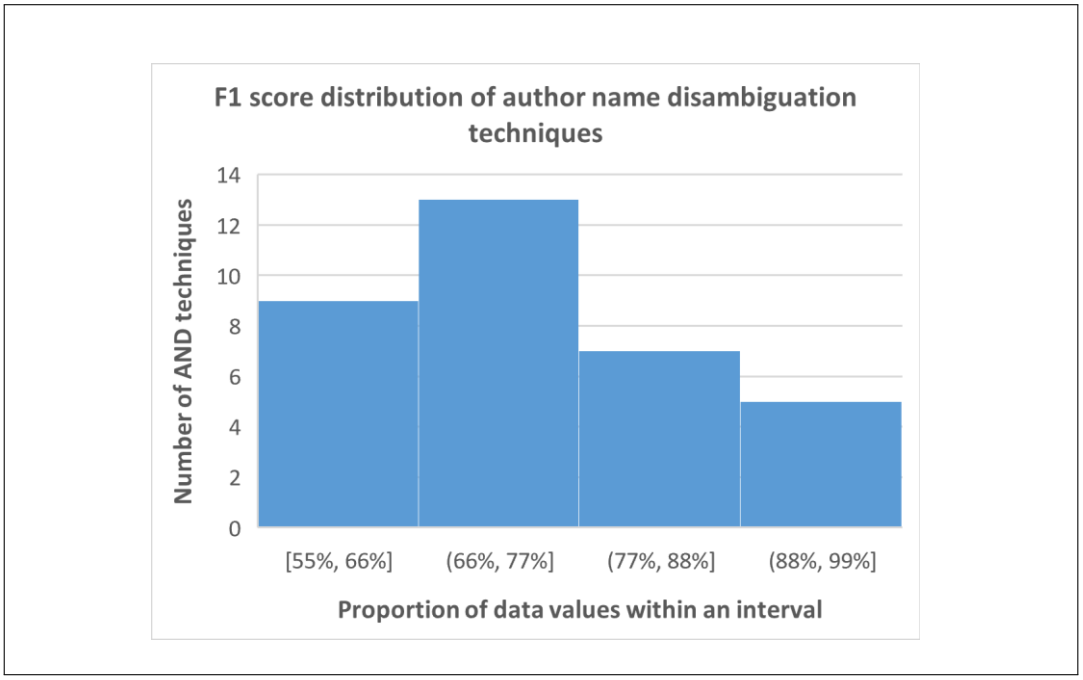

FIGURE 2.20: F1 Score Distribution of the Reviewed Techniques.

However, a few techniques demonstrate F1 scores exceeding 90%. It is important to note that these higher-performing techniques often utilize domain-specific or ethnically-centric datasets, or they are tested on limited datasets. Consequently,

their exceptional results may not be replicable when applied to more diverse, real-world datasets.

The next section gives a comprehensive analysis of the techniques that analyzed features that are used to solve the author name ambiguity problem while making academic authorships. The analysis first criticizes the studies that focused their attention on this aspect and identifies the gaps that need to be addressed in this regard. The analysis is later extended to the reviewed datasets and the problems that need to be addressed in them.

2.2.2 Analysis of Existing AND Studies with Respect to Impactful Features and Datasets

This section critically analyzes the existing AND techniques with respect to features and the datasets used by them. The first subsection encloses details against the features which are used by the techniques as follows:

2.2.2.1 Features

Among the existing AND studies, some have evaluated the effects of features on author name disambiguation results (Table 2.4 in section 2.1). For instance, P. Treeratpituk et al [40] and, D. Vishnyakova et al, [50] explicitly ranked AND features, however, they have only considered a subset of features for this purpose, leaving questions like whether other features are less important? Or, they were not evaluated because of their unavailability in the datasets they used in their studies? Also, some contradictory reports are observed in comparison to their findings in the literature. For example, in Treeratpituk et al [40] proposed feature ranking, author affiliation and co-authors features are ranked low (claiming they are not very useful), whereas similar features are claimed to be very powerful to enhance the academic authorship results by other researchers like J.W. Seol et al [12], J. Zhu et al [18] and D. Vishnyakova et al [50]. Though the studies conducted by P. Treeratpituk et al and, D. Vishnyakova et al explicitly ranked a set of features,

they lack the insight into different better result-producing feature combinations to enhance the results.

It is also observed that few studies explicitly provided some features ranking, whereas, others have implicitly carried out experimentations on their techniques to rank the features. Like, J. Kim et al [1] have implicitly talked about only four features and their combinations impact on their technique results. Similarly, J. W. Seol et al [12], implicitly gave insight into only five features and the impact of their combinations. J. Zhu et al [18] implicitly discussed only three features and their combinations.

Based on the review of techniques with this perspective, Figure 2.21 summarises that out of total nine techniques, how many studies considered a particular feature as important. This means that higher percentage value against a particular feature represents that it is declared a useful feature with more votes, whereas low percentage means otherwise. Similarly, Figure 2.22 is plotted based on the statistics given in Table 2.8. This graph represents the percentage of the 17 useful features (listed in Table 2.5) usage by the reviewed AND techniques. The higher percentage represents that the feature is used by more techniques, whereas the low percentage score represents that the useful feature is used by fewer techniques.

Considering the graphs shown in Figure 2.21 and Figure 2.22, it can be observed that both of them are almost in line with each other. This means that the features that are considered useful by the research community are also being used by researchers to enhance academic authorships.

To sum up the feature analysis, it is concluded that the existing studies that focus on assessing the impact of features on the under-discussing problem lack a complete picture. They have reviewed a subset of features, without talking about many other features. The literature lacks a concise individual feature ranking list, as we have encountered many contradictory reports against the features. Similarly, many features which are ranked high cannot potentially remove ambiguity rather increase it. e.g. Treeratpituk et al [40], ranked author last name at the top, however, it is known that the authors name ambiguity occurs because of the

same names, or due to their name variations. Hence, this feature cannot distinguish distinct authors if their names are identical or have the same name variant i.e. first name initial and last name. Same argument applies to author's middle name which is given rank 2 by the same authors. Besides this factor it is also observed, that the literature lacks knowledge about the feature combinations that can contribute to make better academic authorship, and thus can enhance author name disambiguation.

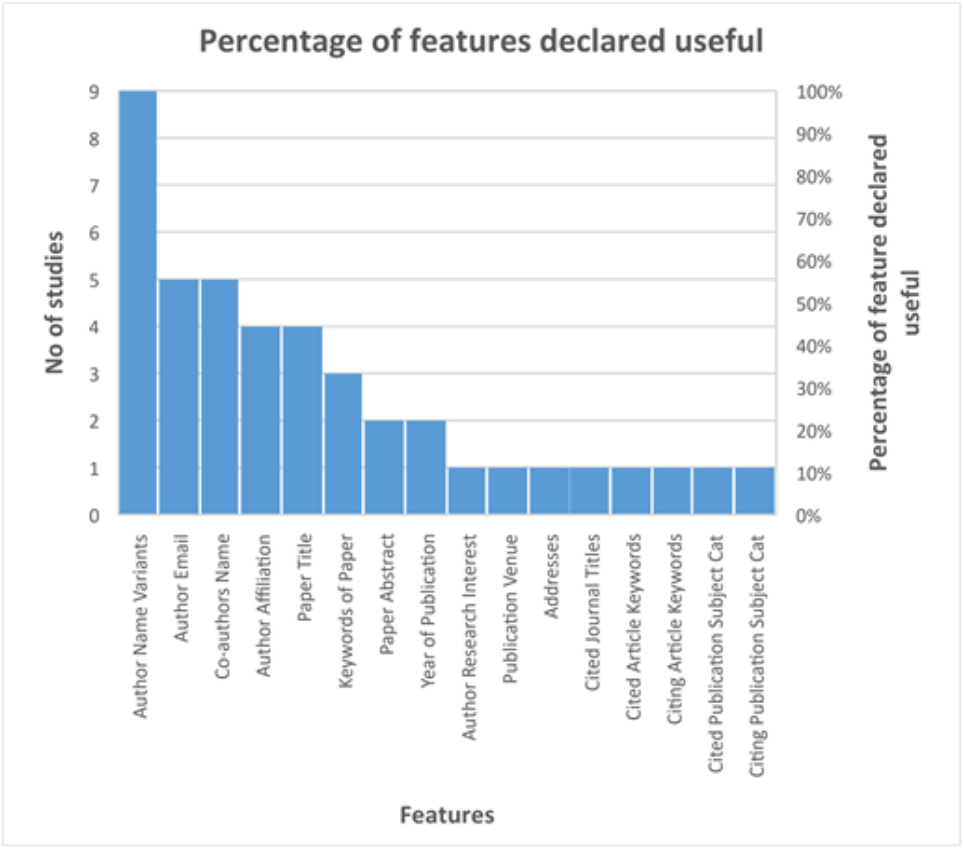

FIGURE 2.21: Percentage of Features Declared Useful by Reviewed Studies

The next section of this study analyzes the publicly available datasets with respect to their labeling strategies, ethnicity of authors, availability of useful features in

them, and the domain(s) of the data in the dataset.

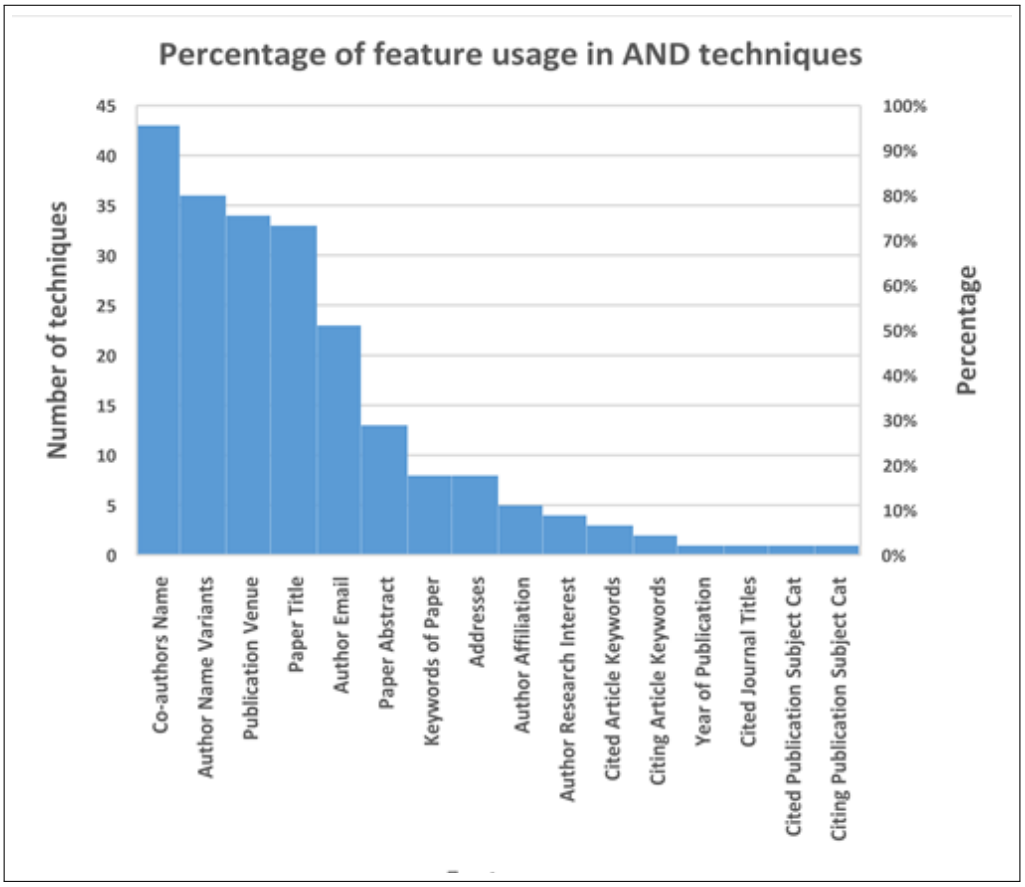

FIGURE 2.22: Percentage of Feature's Usage in AND Techniques

2.2.2.2 Datasets

This section analyzes the reviewed datasets (listed in Table 2.6, section 2.1) as follows.

Datasets Labeling Strategy:

The curation or labeling process of the datasets can be manual, automatic, or semi-automatic. Most of the publicly available AND datasets are hand-labeled, which is a daunting and expensive task. As an alternative, some scholars use list of name pairs that match on specific criteria to automatically generate large-scale

datasets as proposed by A. A. Ferreira et al [22], M. Levin et al [42], and V. I. Torvik et al [86].

Despite the large sizes of the automatically generated datasets, the matching based labeling methods for dataset curation process have some common problems, which are also identified by H. Waqas et al.[15]. For instance: 1) Their matching criteria are rarely verified for accuracy. 2) Their performance relies on the availability of information e.g., co-authors based matching may result in poor performance in the case of sole authorships, or, in the case of small teams. Whereas, in large teams, there is a probability that the co-authors themselves are ambiguous. For ORICID[11] vs DOI[12]-based matching, as done by J. Kim et al [87] and L. Zhang et al [88], they either do not maintain the author's position or rely on schemes for their position identification, which are often not verified due to the large number of instances.

Ethnicity of Authors:

In a recent study conducted by J. Kim et al. [16], it is identified that author names from various ethnic groups (EGs) exhibit distinct patterns of ambiguity. For instance, some ethnic groups might have a higher prevalence of common surnames, making it difficult to distinguish between authors with similar names. Others might use different orderings or combinations of given names and family names, adding to the complexity of accurate name disambiguation. The study also identified that the publicly available datasets for AND, often have a non-uniform distribution of EGs. Due to this uneven distribution the models trained on datasets that do not adequately represent the diversity of EGs may perform well on the majority group but poorly on others, leading to biased results. While diverse ethnicity's introduce varying degrees of ambiguity, the inclusion of all EGs in the datasets is not obligatory. Nevertheless, it is better that if any EGs are included, they should not exhibit a skewed distribution. As the uneven distribution of EGs creates significant issues for the generalization of AND techniques.

[11]It is an alphanumeric code to uniquely identify authors and contributors of scholarly communication

[12]It is an alphanumeric unique code which is used to identify a digital article or document

To conduct the EG analysis of the reviewed datasets, EG tagging is done in this study. For this purpose an EG tag is assigned to an author name instance of a dataset using the author name ethnicity classification database, Ethnea [13]. Based on the EG tagging, the statistics shared in Table 2.6 (section 2.1) show the complete picture of the EGs included in the reviewed datasets. It can be observed that some datasets hold only one specific EG in it. For example BDBComp. Some datasets have a higher percentage of one EG whereas other EGs are skewed. Thus making the entire dataset unbalanced in this respect, which will hinder in developing a fair and effective disambiguation model [1, 15]. Therefore, it can be concluded that the reviewed AND datasets are skewed in this respect [15].

Useful Features Coverage and the Domain of the Data in the Datasets:

Refer to Table 2.7 (section 2.1) for the statistics related to the features availability in the reviewed datasets. It can be clearly observed that the majority of the reviewed datasets do not even cover half of the useful features in them. Similarly, some of the most distinctive features for AND are not even available in 50% of the reviewed datasets. As far as the domain of the publications in the datasets is concerned, it is observed that all of the datasets are domain-specific. For example, Medline and PubMed have only medical-related data. Inspire is based on the Physics domain, and scadZBMATH includes data from the Maths domain, Whereas, other datasets have only CS related publications in them. However, it is greatly emphasized in the literature [15, 16] that different domains introduce diverse scenarios and patterns in the author name ambiguity problem. The lack of such diverse scenarios and the absence of useful features in data brings a challenge to test and develop generalized solutions for the under-discussing problem.

This concludes that the publicly available AND datasets lack useful feature coverage in them, they have skewed ethnic distribution of ambiguous authors and they hold data that are limited to a specific domain.

[13] Ethnea is developed by Torvik et al [89], (2016), which is a collection of more than 9 million author name instances that are tagged one of 26 EG classes based on the name's association with national-level geo-locations. This is a similar EG tagging strategy that is adopted by J. Kim et al, [16]

2.2.3 Conclusion and Problem Statement

From the analysis of the literature review, several key problems are identified with existing Author Name Disambiguation (AND) techniques:

Techniques that rely on word similarities for authorship disambiguation fail to effectively handle cases where authors work in the same domain, like the techniques proposed by R.G.Cota et al[20], A.A.Ferreira et al, [22], D.Shin et al [17]. This limitation arises because such techniques cannot distinguish between different authors whose research topics or keywords significantly overlap. Techniques utilizing embedding models trained on specific training sets struggle with new data, particularly for authors with limited publications or those working in diverse domains, like the technique proposed by K. Pooja et al [8], [71], Peng et al [13]. These models often lack generalizability beyond their training datasets. Similarly techniques that employ pre-trained embedding models may produce sparse vectors, leading to compromised recall and precision. The embeddings may not adequately capture the nuances required for accurate author disambiguation.[90]. Whereas techniques that rely solely on overlapping co-authors or keywords suffer from low recall and often face precision issues. Different authors might share common co-author names or keywords, leading to incorrect disambiguation, as proposed by A.A.Ferreira et al [22], B. Xiong et al[64].

Furthermore, the literature reveals gaps in feature rankings and combinations which reveals that there is no consolidated list of feature rankings, making it challenging to identify which features contribute most effectively to author name disambiguation. Whereas either the ranked features are domain specific or are prone to adding false positives, thereby exacerbating the ambiguity rather than resolving it like the rankings proposed by P.Treeratpituk [40], M. Levin et al [42], D. Vishnyakova [50]. Also, there is a lack of comprehensive knowledge about feature combinations that could enhance the overall F1 score.

From the key limitations identified from the literature review, we can conclude as follows:

1. The analysis of the AND techniques points toward the need to improve the overall academic authorship results [34].

   The study indicates that when a technique enhances precision, it often leads to a reduction in recall, and when recall is improved, precision tends to decrease, ultimately affecting the overall F1 score. This gap is also highlighted by different researchers like H. Waqas et al [34], Pooja et al [71].

2. To improve the AND technique's result, the literature lacks knowledge of feature combinations that can positively impact its output [34].

3. AND datasets are not sufficiently feature enriched (also highlighted by H. Waqas et al [15]), they have unbalanced authors' EGs (J. Kim et al and H. Waqas et al., [15, 16]) and lack data of multi-domains. The literature analysis points that different ethnicities bring complexity to the author name ambiguity problem and multi-domains introduce diverse scenarios and patterns in the data which elevates the author name ambiguity problem [15]. Hence, the presence of these aspects in the dataset will introduce reasonable complexity in them which will later become the basis for developing and testing generalized AND techniques.

From the key limitations identified from the literature review, and to provide a clear understanding of the practical impact of these limitations: the following Research problem is devised:

"Existing AND techniques generally have low precision, recall, or both, which affects their overall results. If a technique manages to raise precision, its recall often gets compromised and if the recall is achieved its precision generally becomes low, which subsequently results in low F1 score."

The following chapter focuses on filling the gap of lack of knowledge of feature combinations which can give better academic authorships, along with individual feature rankings.

Chapter 3

Features Combinations Impact

Due to the partial availability of knowledge about the impact of feature combinations on the overall result of the AND technique, the focus of this chapter is to answer a part of the research question 1 which is:

1. How to devise an AND technique that can perform academic authorships with improved results, without compromising its precision? To devise such an AND technique: **What features combinations produce higher precision and recall in order to get better AND results?**

3.1 Introduction

To study the impact of features on the authorship results, and identify better results producing feature combinations along with individual feature rankings, this study use a wrapper based Sequential Forward feature Selection (SFS) technique. SFS is applied to three distinct candidate features lists, where, two of these lists are obtained from literature, while the third list is generated through Principal Component Analysis (PCA) using Pearson's correlation coefficient [31, 32]. The entire process identifies a list of different feature combinations (ranging from single to multiple) that achieve better authorship results.

The motivation behind using SFS for this purpose is to filter out feature combinations without increasing the complexity of architecture design. Whereas, its common problem of exhaustive searching in the feature space is catered by initially devising a set of candidate features from the literature. Similarly, the motivation for using PCA in this study is to filter out features with the most variations from a list of features set, without increasing the complexity of architecture design, hyperparameter tuning, and large amounts of data requirements as in the case of neural networks [91].

Rest of the chapter is organized as follows: Section 3.2 formally defines how the features are assessed. Section 3.3 describes the approach adopted to make AND feature ranking, which is followed by the experimental details. Section 3.4 highlights the results, which is followed by the discussion section (3.5) section. A dedicated section is introduced before the chapter summary to emphasize on the novelty of this work.

3.2 Formal Definition

Given a set of papers $P = \{p_1, p_2, \ldots, p_k\}$, where k is the number of papers, the system aims to group all papers authored by the same author into clusters using one feature at a time, resulting in a set of clusters $C = \{cs_1, cs_2, \ldots, cs_n\}$.

For any paper $p_j \in P$, the feature set is represented as $F_{extract\_features}(p_j) = \{f_1, f_2, f_3, \ldots, f_m\}$, where $1 \leq m \leq 17$ (useful features). The goal is to identify individual and combinations of features that yield the highest pairwise F1 score ($pF1$) for papers authored by the same author.

3.2.1 Feature Selection Criteria

1. **Clustering Using Single Features**:

    - For each feature f_i in the feature set $F_{extract\_features}(p_j)$, cluster the papers P to form clusters C_{f_i}, where $C_{f_i} \in C$.

- Compute the pairwise F1 score ($pF1_{f_i}$) for the clusters C_{f_i}.

2. **Feature Combination and Ranking**:

   - For any combination of features $F' \subseteq F_{extract\_features}(p_j)$, cluster the papers P to form clusters $C_{F'} \in C$.
   - Compute the pairwise F1 score ($pF1_{F'}$) for the clusters $C_{F'} \in C$.
   - Rank the features and feature combinations based on their $pF1$ scores.

A feature or combination of features is considered more useful and is ranked higher if it results in a higher $pF1$ compared to others.

3.3 Methodology of Feature Ranking

The methodology adopted to propose our feature ranking and identify better results producing feature combinations we have devised the following novel methodology (as to the best of our knowledge not previously adopted in this domain):

1. First, we have identified candidate features which will compete with each other for this purpose. We have proposed three candidate sets with preliminary ranks following three different schemes. The granular details of this process is given under subsection 3.3.2.

2. After the first step a SFS approach is adopted to test each feature's capability to resolve author name ambiguity both individually and also in combination with other features. The feature combination is based on the preliminary feature rank given to them in the beginning of the process i.e. as described in first step.

3. After running the experimentation the initial ranks are updated which are the final rank list.

4. To conduct feature ranking experimentation, the participating datasets are chosen based on the availability of the highest number of features availability in them. The detailed process is covered under section 3.3.3.

5. To evaluate the features contribution to solve AND, Multi-layer Hierarchical Clustering technique is used. The details are listed under section 3.3.4, and in Chapter 4.

3.3.1 Workflow of Feature Ranking:

The workflow to identify better results producing feature combinations and individual feature rankings, is divided into three major processes as shown in Figure 3.1.

Process 1:

It is involved in producing three candidate features lists to reduce the initial feature search space for SFS. Two lists are generated from the literature survey, whereas one list is produced by following PCA based feature reduction. The candidate features are identified with an intermediate rank based on a specific criteria, such that these ranks are later readjusted.

Process 2:

It is related to dataset identification and selection, such that they cover majority of the candidate features to test and evaluate their contributions. Two datasets are selected following this criteria, whereas a third dataset "CustAND" is curated (Chapter 5) for this purpose (as the publicly available datasets have limited features. Details are already discussed in Chapter 2).

Process 3:

It is related to the identification of better result producing feature combinations and individual feature ranking. The following subsections explain each process of the workflow in detail.

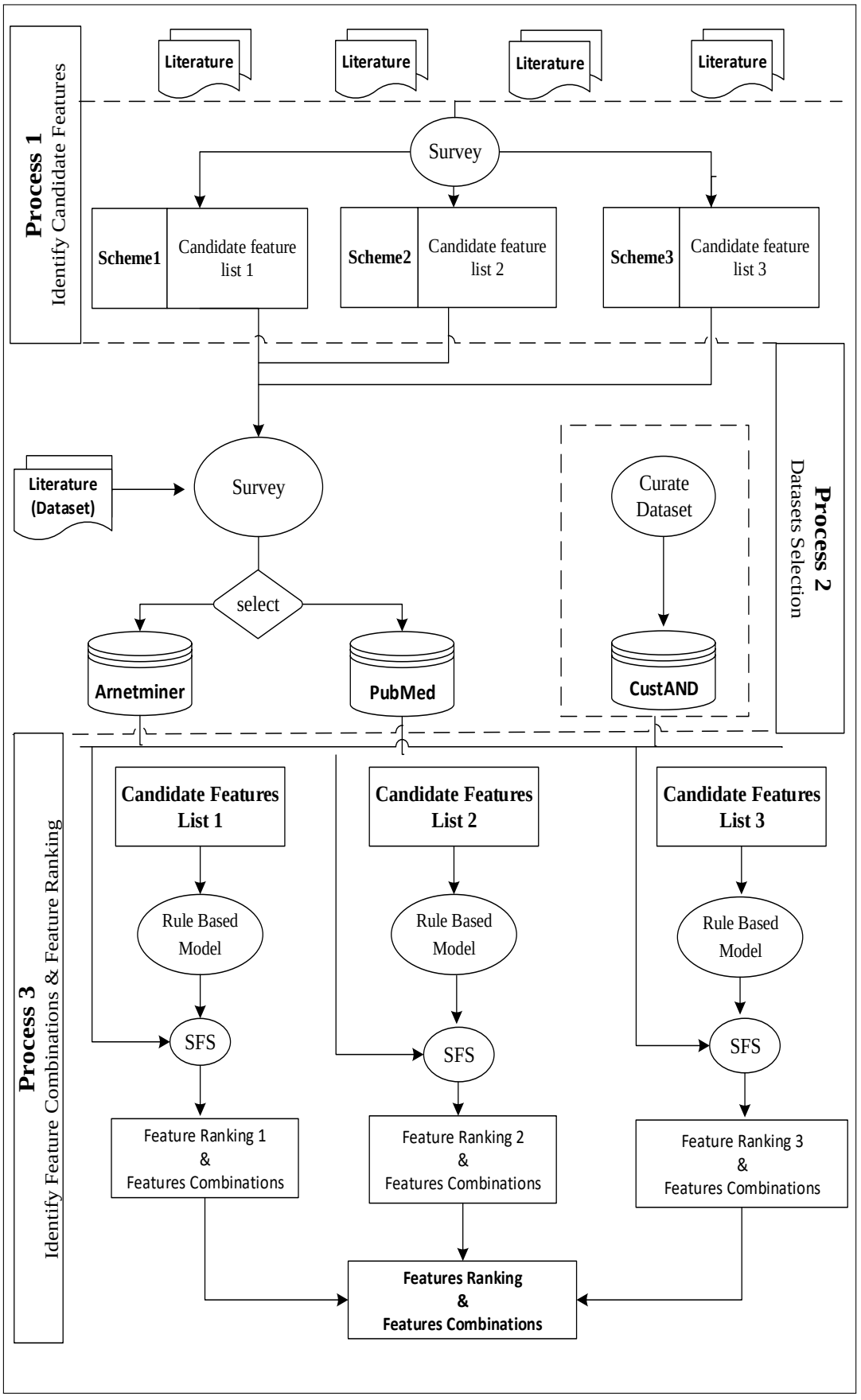

FIGURE 3.1: Overall Workflow to Identify Better Results Producing Feature(s) Combinations and Individual Feature Ranking

3.3.2 Process 1

In this process, three candidate feature sets are formulated using three schemes. **Scheme 1** is based on the literature review of studies assessing the usefulness of AND features as comprehended in Table 2.5 of chapter 2. **Scheme 2** is based on the literature review of AND techniques with respect to the feature usage in them as comprehended in Table 2.8 of chapter 2. **Scheme 3** is based on PCA based feature reduction of useful features (listed in Table 2.5), comprehended in Table 3.1.

3.3.2.1 Candidate Features Based on Scheme 1

Using statistics from Table 2.4, intermediate ranks are assigned to the identified useful features (Table 2.5). The ranks are computed such that the frequency of the features assessed in literature is considered the weight (w), times the average of the features declared to be useful. Therefore, the intermediate rank can be computed using the formula given in equation (3.1 and 3.2), whereas the ranks given to the features using the proposed equations are validated using the experimental results given in the upcoming section.

$$u = d/a \tag{3.1}$$

$$Rank = f * u \tag{3.2}$$

The ranks are prioritized in ascending order, and :

f = number of times a feature is assessed in the literature (reviewed in this study).

a = total number of studies included in the literature survey to study the AND features impact on the authorship result assessment i.e., 9.

d = number of times a feature is declared useful in the literature.

u = usefulness of a feature with respect to the total studies.

The ranked[1] list is shown in Table 3.2.

3.3.2.2 Candidate Features Based on Scheme 2

Candidate features ranking using scheme 2 is based on the literature review of AND techniques with respect to the feature usage in them. For this purpose, the statistics of Table 2.8 are comprehended in Table 3.2, such that the highest ranked feature corresponds to the feature being used most frequently by the existing AND techniques.

3.3.2.3 Candidate Features Based on Scheme 3

PCA Based Features Evaluation and Reduction

For PCA based feature analysis and reduction, set of sample papers are taken from Arnetminer, CustAND and PubMed datasets, which are clustered manually by observing set of features (one by one). Papers are observed pairwise and scores ranging from 0 to 0.9 are assigned to features such that their cumulative score is equal to 1. Scores are assigned to a feature or set of features that will help to correctly identify the author of the paper (group), such that the actual label or author id is hidden from the observer. A feature is scored highest if it can identify the correct cluster independently, else, the score is distributed among the next supporting feature. An observed feature is scored low if the observer cannot decide the correct author group with surety while using it individually. A score greater than 0.1 is assigned to a feature which is observed in combination with another feature if and only if it is helpful in identifying the correct author group. Any feature is given a score zero if it is not used in the disambiguation process, whereas 0.1 score is given to a feature which is observed but cannot distinguish the author group. For sample paper selections 167 random numbers are generated between a range starting from zero to the number of papers having no null values

[1]The results of intermediate feature ranking based on the procedure explained above are given in Table A.1 of Appendix A.

against all features (for each dataset). Publications against the index numbers present in the random number list is selected from each dataset. Therefore a total of 501 samples are collected. After manual score assignments, PCA is applied to these observations, to obtain the reduced features set.

Generally in PCA[2] the projection directions that capture the most variance are considered, where the directions with the most variance are the ones with most inertia [92]. Keeping in view this concept and avoid misleading results some feature pruning is done based on heuristics. All features with zero scores against all instances are removed, i.e., author name variants, cited papers titles, cited journal names, paper language, cited articles co-authors, etc. Also, all features that are majorly assigned 0 scores and rarely with scores less than 0.3 are removed, e.g., publication date, to avoid the generation of higher variance and eigenvalues as compared to frequently used features with higher score values and sparse 0's, like, author affiliation, author email, paper title, etc. Author full name feature is also discarded from the analysis as this feature is already ambiguous and no one can disambiguate any publication solely using this feature if it is not distinct. Further, in majority of the observations this feature is assigned a score equivalent to 0.1 and rarely 0.9 (only if the author's full name is unique which is rare in real-life scenarios).

Based on the experimental output, it can be said that the participating features within this PCA-based analysis are as reported in Table 3.1. These reduced features will further participate in the individual feature rankings as compared to all 17 useful features that are identified through literature.

[2]Table A.12 of Appendix A, shows the overall statistics of the features and the data used for this experiment. Table A.13 of Appendix A reports the individual and cumulative variability with respect to the eigen values of the factors and Figure B.1 (Appendix B) shows the eigen values of the features versus their cumulative variability percentage. Whereas, Table A.14 of Appendix A shows the correlations between features and their factors. The chart in Figure B.2 (Appendix B) shows the data projection using two PCA axis representations in two-dimensional space. Whereas, details regarding attaining the reduced feature list is given in Appendix B, before Figure B.1.

TABLE 3.1: Candidate Features Based on Scheme 3.

Candidate Features
Co-authors Names
Author Affiliation
Author Email
Paper Venue
Paper Title

TABLE 3.2: Candidate Features Based on Scheme 1 and 2.

Candidate Features	**Scheme1 Rank#**	**Scheme2 Rank#**
Author Name Variants	1	1
Co-authors name	2	2
Paper Ttile	3	4
Author Email	4	9
Author Affiliation	5	5
Paper Keywords	6	7
Paper Abstract	7	6
Publication Year	7	8
Publication venue	8	3
Author research area	9	9

Journal Title	-	10
Journal Language	9	12
Address	10	12
Citing Article Keywords	11	11
Cited Publication Subject Category, Citing Publishing Subject Category	11	12

3.3.3 Process 2

In this process, three datasets are selected to evaluate the impact of features. The basic selection criteria of the datasets are:

1. Maximum number of useful features availability in them (refer to Chapter 2, Table 2.7, and Section 2.2.2.2 for datasets review, analysis, and findings related to this criteria).
2. Most commonly used dataset by the studies that have assessed the AND features, and by most of the existing AND techniques.

Following selection criteria 1, **PubMed** [74], is selected, which holds medicine-related scholarly data, covering eleven features. It is preferred over Medline for this purpose even though both cover eleven features, because of the unavailability of the author's full name feature in Medline.

The second dataset is **Arnetminer** [18, 25][3], which is selected based on the above mentioned criteria 2, even though it covers only six useful features, holds Computer Science intensive scholarly data, and is curated for feature-scarce scenarios.

The third dataset is **CustAND** (refer to chapter 5 for dataset details), which is curated specifically to study the impact of features and their combinations on the

[3] https://www.aminer.org/disambiguation

authorship results. CustAND is curated based on the literature analysis and gaps in the existing publicly available datasets (refer to chapter 2, section 2.2.2.2).

3.3.4 Process 3

Process 3 is the rule based model which takes input from process 1 and 2 in the form of three candidate feature lists and datasets respectively. The rule based model evaluate the features impact on the authorship results based on sequential forward feature selection process. Ultimately giving two outputs, 1) Single feature ranking based on pF1 scores. 2) Better pF1 score producing feature combinations.

The working of the rule based model in process 3 is given in detail in the next section, whereas the rules, algorithms, and workflow is covered in Chapter 4. Moreover, the evaluation metrics to assess the authorship results are listed in Chapter 1, Section 1.7.1.

3.3.4.1 Rule Based Model

Features serve as clues to the rule-based predictor model that takes as input two papers and decides whether they should share the same group or not based on a particular feature.

The rule based model takes the three candidate lists one at a time, and groups the input papers into ambiguous blocks with respect to the author's short name. After this step the system tries to merge two groups based on the candidate features one by one. After each merge, the system repeats the process until no further merging is possible, and reports the results. The system then takes the next feature and repeats the process. In case a feature shares the same rank with another feature(s), it calculates and reports the performance of features by considering their combinations separately. For example, if feature C and D both shares rank #3, two separate combinations of features are considered i.e. A (feature used at rank #1), B (feature used at rank #2), **C** and A, B, **D**.

To ensure the model's robustness and to identify the most influential features, an ablation process is employed. In this process, each feature is systematically removed from the candidate list one at a time, and the system's performance is reassessed. This allows for an understanding of the contribution of each feature to the overall performance. By comparing the results before and after the removal of each feature, the model identifies which features are most critical to the clustering task and which may be redundant or less influential.

The rule based model groups the features in two ways: 1) Structure aware features. 2) Global features. The details of these two groups are given as follows:

Structure Aware Features

Structure-aware features are fixed sets of attributes related to publications and authors (for instance, author email, author affiliation, co-authors, paper publishing venue, and paper publishing date). Considering the values of these features, it is more likely that the papers belong to the same author if they contain the same keywords. Like, it is more probable that common terms are present in the affiliation or emails of the author in question [4], as well as some common co-authors' names with whom the researcher often collaborates. Therefore, co-authors' names, author's affiliations, authors' emails, paper publishing venue, and publishing date are grouped as structure-aware features, in the rule-based model.

The similarity measures used for structure aware features comparison by the proposed rule based model varies per feature, which are given as under:

1. Cosine similarity[5]: This similarity measure is used to compare author's emails, author's affiliations, and paper venues, as the they are likely to be represented using same keywords.

2. Binary similarity: This measure is used to find common co-authors, as it is more likely that the papers authored by multiple common co-authors belong to the same author.

[4] author who is being considered for disambiguation

[5] Cosine similarity is a measure of similarity used between two non-zero vectors that measures the cosine of the angle between them.

3. Coauthor name fragment checking algorithm: This algorithm checks coauthor name fragments in case the binary similarity fails. This will cater the cases when the author's names have spelling mistakes, or have different name fragments. For example, "M. Abdul Qadir" and "Muhammad Abdul Qadir" etc. Refer to Chapter 4, Algorithm 4, for details.

Global Features

Global features are extracted from the textual features of the publication, including, paper title, abstracts, and keywords, using the proposed word2vec inspired trained model named **Research2vec**[6]. Research2vec allows the system to consider the semantic similarity of terms instead of merely looking for common keywords within these features. Research2vec is explicitly trained on scholarly data, making it more adept at predicting word contexts compared to pre-trained Word2Vec models, which are typically trained on generic or domain-specific corpora. The Research2vec model is trained using the Continuous Bag of Words (CBOW[7]) architecture with 300 dimensions, with a context window size of 4, on a corpus containing 3.3 million abstracts of research papers published on arXiv[8]. The raw data is available on Kaggle[9] and can be downloaded from the given link[10]. In contrast, pre-trained Word2Vec models, such as those trained on the Wikipedia corpus, also use the CBOW architecture with 300 dimensions but typically employ a context window size of 5. These models are trained on a large, diverse set of Wikipedia articles, making them versatile for general applications but less specialized for academic research contexts.

Before proceeding to discuss the results, consider the following real example to understand the entire process:

[6]Research2vec model can be downloaded from: https://drive.google.com/file/d/ 1JHwPIC-jAWXfut6o86kILSzP2BpfQYk2G/view?usp=sharing

[7]CBOW architecture of a model learns the context of the words and tries to predict words which are contextually similar.

[8]arXiv is a free distribution service and an open access repository for scholarly articles in different fields.

[9]Kaggle is the world's largest data science community with tools and resources to help achieve data science-related goals.

[10]https://www.kaggle.com/Cornell-University/arxiv.

Walk-through the Rule-Based Model: An Example

Consider the instances in an author block "M Qadir", where the respective feature values are given in Appendix A[11]. As mentioned above, the rule based model takes the candidate features lists one by one and assess the features contributions. Consider, for instance, the rule based model considers the candidate scheme 1, which enlists co-authors names as a top-raked feature. In the example, the first and second instance co-authors (Syed Zubair Ahmad, Mohammad Saeed Akbar) and (Muhammad Fahad, Muhammad Wajahat Noshairwan, Nadeem Iftikhar) are used to find common co-authors, using the Algorithm 4, given in Chapter 4. If the common co-author count is more than 1 excluding the author in question, the instances are merged, else, the model will try to match the first with the next instance, until no further merging is possible (only using the current evaluated feature i.e. co-authors name). For example: comparing instance one co-authors (Syed Zubair Ahmad, Mohammad Saeed Akbar) with the third instance (Umar Farooq, Antoine Nongaillard, Yacine Ouzrout). Then the next instance, (Syed Zubair Ahmad, Mohammad Saeed Akbar) with (Umar Farooq, Antoine Nongaillard, Yacine Ouzrout). In the given example, instances three and four are merged based on the rules specified using the co-authors feature. Therefore, after the entire process is complete, the precision, recall, and F1 scores are computed against a given name block.

Now the process is repeated using the next candidate feature on the newly grouped data, as the number of instances will be 3, and the feature that will be used to merge the data is the next candidate feature which is the paper title. Therefore, now instance one with paper title (High Speed Scalable Mobility Management Architecture over Infrastructural WLAN) is compared with instance two (DKP-OM: A Semantic Based Ontology Merger), such that global features are extracted from both titles using the proposed Research2Vec model, and then compared. Two instances are grouped together if the cosine distance between the global features of the two instances are within a reasonable threshold value. The algorithmic details and threshold value selection criteria is discussed in Chapter 4. The entire

[11] Table A.29

process using this feature is repeated until no further merging is possible. A similar procedure is adopted for the other candidate lists too.

The next section reports the results of the feature ranking workflow.

3.4 Results

The results reported in this section are based on the SFS technique when applied to three candidate feature lists (listed in Table 3.1 and 3.2), one by one, using three datasets (Arnetminer, CustAND, PubMed). All of the results are comprehended in two ways. 1) By identifying set of feature combinations which give highest pF1 scores. 2) By identifying individual feature rankings based on highest pF1 scores.

3.4.1 Features Combinations Based on pF1 Scores

The detailed results of SFS technique when applied on candidate features (scheme1) are given in Appendix A[12] which are comprehended in Table 3.3 as follows.

Table 3.3: Highest pF1 Based Feature Combinations.

Dataset	Features Combinations	pP	pR	pF1
Arnetminer	co-authors, author affiliation	99%	91%	95%
	co-authors, author affiliation, paper title	98%	92%	95%
	co-authors, venue, author affiliation	89%	94%	91%
CustAND	**author email, author affiliation**	**100%**	96%	**98%**
	co-authors, author email, author affiliation	97%	99%	98%
	co-authors, author email	97%	96%	96%
	co-authors, venue, author affiliation	94%	96%	95%
PubMed	co-authors, author email	83%	73%	78%

[12]Table A.2, A.3, and A.4. Similarly, Table A.5, A.6 and A.7 of Appendix A covers the detailed results when SFS is applied on scheme2 based candidate features. Moreover, Tables A.8, A.9 and A.10 of Appendix A cover the detailed results of candidate features (scheme3). In these tables, highest Δ pF1 achieving feature combinations are given in bold. The identified combinations from Table A.2 till A.10

author affiliation, author email	89%	69%	77%

3.4.2 Features Ranking Based on pF1 Scores

To find the individual feature ranking, SFS approach is applied to 5 features which are author email, author affiliation, co-authors, paper title, and paper venue. The features are ranked based on pF1 results[13]. Whereas, the ranked features are listed in Table 3.4, using three datasets.

TABLE 3.4: Individual Feature Rankings Based on pF1 Scores (NA means the feature is absent in the dataset).

	pF1 attained by individual features with respect to datasets				
Datasets	**email**	**affiliation**	**co-authors**	**paper title**	**paper venue**
Arnetminer	NA	86%	85%	73%	15%
Rank#	-	1	2	3	4
CustAND	90%	83%	65%	62%	17%
Rank#	1	2	3	4	5
PubMed	59%	74%	45%	1%	19%
Rank#	2	1	3	5	4

3.5 Analysis

3.5.1 Proposed Versus Existing Feature Rankings

As discussed in Chapter 2, very few existing techniques have explicitly given feature rankings. Table 3.5, lists the existing feature rankings along with the ones proposed in this chapter.

[13]Detailed results of the experiment are given in Table A.11 of Appendix A

The proposed feature rankings 1 and 2 are the same except that the author email feature is not ranked when the feature combinations are assessed using the Arnetminer dataset. This is because it is absent and its impact on the authorship results cannot be evaluated. However, in the proposed feature ranking 3, author affiliation feature is given a higher rank as compared to author email, because the proportion of records with author affiliation feature is 94% as compared to author email that is 47.2%, in PubMed (the dataset which is used to assess the impact of the feature on the authorship results) [74]. Though the precision score of author email is better than author affiliation feature but the recall is low which affects the pF1 score[14].

As far as the existing feature ranking is concerned, Treeratpituk et al [40], ranked author last name at the top, however, it is known that the authors name ambiguity occurs because of the same names, or due to their name variations. Hence, this feature cannot distinguish distinct authors if their names are identical or have the same name variant i.e. first name initial and last name. Same argument applies to author's middle name which is given rank #2 by the authors. Though this feature can be helpful in case when co-authors feature is used to identify common co-authors, however, its use case is limited. The third ranked feature by the authors is author affiliation, but, in these experiments, it has proved to be one of the top features with high pF1 score, using all three datasets. This can be because in the dataset (Medline) used by the authors, only 61% of the records have author affiliation feature [74]. Whereas in the proposed ranking 3 (using PubMed), this proportion constitutes around 94% [74].

Levin et al [42] ranked the citation-related features as the top useful features to make near to correct authorships. However, it is observed that any keywords-based feature (e.g. paper titles, abstracts, keywords, mesh terms, citation titles, etc) if used to identify distinct authors, often leads to high false positives. The results[15] show that features like paper abstracts, paper keywords, and paper titles do not contribute to the overall authorship results. This is because it is rare that the

[14] as shown in Table A.10 of Appendix A
[15] given in Appendix A, Table A.6

same author uses exact keywords in his papers. It is also possible that different authors use the same keywords in their research papers which are cited by someone else, giving a false perception about the authorship. Moreover, the email feature is ranked low by this study, however, the dataset used by the authors cannot be analyzed for this feature because it is not publicly available at the time of this writing.

Vishnakova et al [50], ranked certain other features than the known citation features. e.g. semantic types, ambiguity scores, etc, in their study. The authors ranked Journal descriptors as the top rated feature to disambiguate distinct authors. However, it is professed that it is highly probable that distinct authors can publish in the same journal, and using this feature can lead to high false positives. A word graph which is shown in Chapter 4, Figure 4.2, shows the case where different authors with same names have published their papers in the same venues (a similar feature as journal descriptor), due to which the overall authorship results become low. Moreover, the authors have identified semantics types as the second ranked feature, however, based on the experiments, it is observed that merging two groups based on their title, abstract, keywords, etc., using semantic similarity, in the early phase will pollute the groups with false positives, which will replicate this behavior, ultimately resulting in lowering[16] the pF1 score. Similarly, author's first name which is given rank #5 by the study, is itself the basic culprit of causing ambiguity, therefore it cannot be a good feature to impact authorship results in a positive way. Though the authors have ranked type of organization, country (are part of affiliation), author affiliation and email, but at a low level. This is possible because the proportion of these feature values are less in Medline (used by the study), i.e. 61% affiliation, 6% email [74]. Therefore, based on the experimental results and in comparison to existing studies, it is seen that the proposed features and their rankings should be considered while making an AND technique. These findings and analysis are used by us to design an AND framework, whose results show a positive gain in the overall pF1 scores (details are covered in Chapter 4), in comparison to other such techniques.

[16]Shown in Appendix A, Table A.2...till A.10.

TABLE 3.5: Comparison of Proposed Feature Ranking.

References	**Features Ranking**	**Dataset Used**
Proposed ranking 1	1) Author email 2) Author affiliation 3) Co-authors name 4) Paper title 5) Paper venue	CustAND
Proposed ranking 2	1) Author affiliation 2) Co-authors name 3) Paper title 4) Paper venue	Arnetminer
Proposed ranking 3	1) Author affiliation 2) Author email 3) Co-authors name 4) Paper venue 5) Paper title	PubMed
[40]	1) Author last name (idf), 2) Author middle name, 3) Affiliation (tfidf), 4) Journal year, 5) Affiliation (softtfidf), 6) Mesh shared (idf)	Medline
[42]	1) Citing keywords, 2) Cited keywords, 3) Citing subject cat, 4) Addresses, 5) Cited subject cat, 6) Email, 7) Language, 8) Cited journal titles, 9) Author name initials	WoS

[50]	1) Journal descriptors, 2) Semantic types, 3) co-authors, 4) Ambiguity score, 5) First name, 6) Last name length, 7) Years difference, 8) City, 9) Type of organization, 10) Language, 11) Country, 12) Initials, 13) Affiliations, 14) Email	Medline

3.5.2 Feature Combinations with Highest pF1 Scores

This section covers an analysis of the feature combinations that give the highest pF1 scores, as per the experimental results given in the previous section.

To the best of the knowledge, only few existing studies have worked on ranking AND features, whereas majority of these studies lack the information regarding feature combinations which are less prone to add false positives while making authorships in case of ambiguous author names.

The bar graph shown in Figure 3.2, shows the identified combinations based on the results given in the previous section. The graph shows that author's email and affiliation features make precise groups with the highest pF1 scores. All other combinations result in the inclusion of false positives, lowering the precision and pF1 scores, as, in the case of co-authors feature in combination with author email and author affiliation. Similarly, co-authors with author affiliation, or co-authors with author email feature. Same is the case experienced with paper venue and paper title features.

Therefore, it can be concluded that author email and author affiliation feature combination is the highest pF1 producing combination, whereas other features can be used in case of their in-availability or their sparse values (to cater the recall issues), but this will be at the cost of precision scores.

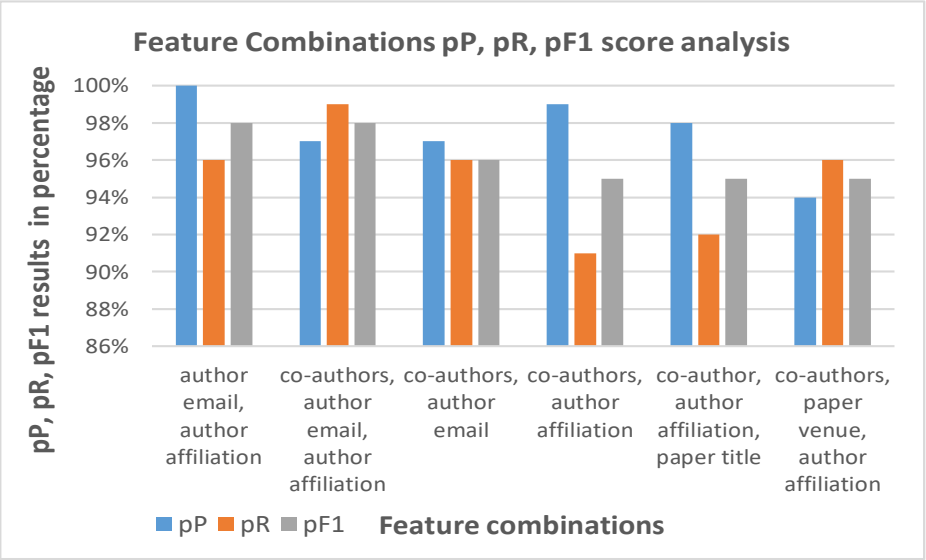

FIGURE 3.2: Feature Combinations with High pF1 Scores.

3.6 Novelty of Feature Ranking Scheme

The development and refinement of feature ranking techniques play an important role in enhancing the performance of Author Name Disambiguation (AND) techniques. This chapter introduces a novel feature ranking methodology that identifies gaps in the literature and optimizes the feature selection process, which was not previously done. The multifaceted contributions based on the novel methodology are outlined below:

1. **Initial Feature Ranking Based on Literature Review**: A literature review-based approach is adopted to assign initial ranks to candidate features and subsequently refinement of these rankings are done through experimental validations. This iterative process ensures a comprehensive and informed selection of features at the end.

2. **Optimized Feature Combinations**: The adopted methodology identifies feature combinations that give better precision, recall, and F1 scores, thereby will help in enhancing the overall accuracy of AND techniques.

3. **Experimental Validation**: To ensure the validity of the proposed feature rankings, individual feature rankings are validated experimentally across multiple datasets. This process rectifies the risk of conflicting results, often observed in previous studies, ensuring the consistency and reliability of our rankings.

4. **Comparative Analysis with Existing Rankings**: A critical comparison of the proposed feature ranking with existing ones, offers a concise perspective to the research community.

3.7 Chapter Summary

This chapter encloses details regarding the study of the impact of features and their combinations on the authorship results. The outcome of the chapter is a list of better result-producing feature combinations, as well as a list of individual feature rankings based on the pF1 scores. To achieve the outcome, the chapter encloses the details of the proposed wrapper-based technique (SFS), which is adopted to evaluate three candidate feature sets that are identified from the literature.

The experiments conducted using the proposed methodology show that the author email feature has the highest rank among the individual features, with a pF1 score of 90%. Following closely, the author affiliation feature is ranked second with a pF1 score of 86%. The co-authors name feature is ranked third with a pF1 score of 85%. The paper title feature is ranked fourth, with a pF1 score of 73%. On the other hand, the paper venue feature is ranked lowest[17] due to its significantly low pF1 score.

As far as the feature combinations which gives better pF1 scores are concerned, the experiments show that author email and author affiliation give highest pF1 scores. Whereas inclusion of any other feature's lead to inclusion of false positives in the groups. Therefor, the findings of this chapter suggests that the identified feature(s)

[17]Appendix A, Table A.11 shows the individual feature rankings based on overall authorship scores.

and combinations can be helpful to formulate simple, scalable, and precise author name ambiguity resolving techniques without the need to exhaustively look out for better results producing feature combinations or using any feature combinations merely because of their availability without considering its impact on the results.

The next chapter of the study discuss the details of the proposed AND technique which will make use of the outcome of this chapter, that is, the ranked AND features and better result-producing feature combinations, along with a simple and improved predictor(s) methodology to achieve better authorship results (in the form of better F1 scores) as compared to the similar existing techniques.

Chapter 4

A Clustering Approach for Author Name Disambiguation

The focus of this chapter is to answer research question 1 whose subpart *"To devise such an AND technique: what feature combinations produce higher precision and recall in order to get better AND results?"* has already been answered in the previous chapter i.e. chapter 3. Therefore, this chapter is a contribution to answer the following part (given in bold) of it:

1. **How to devise an AND technique that can perform academic authorships with improved results, without compromising its precision?** *To devise such an AND technique: what features combinations produce higher precision and recall in order to get better AND results?*

The rest of the chapter is organized as follows. In the first section, the proposed approach to enhance the academic authorships is described formally. This is followed by the proposed approach discussed under "Multilayer Heuristic Based Clustering Framework", which describes the approach in detail. The "Experimental Results" section presents the experimental details, baseline approaches, experimental setup, datasets used to evaluate the proposed technique along with the detailed results. The "Analysis" section highlights the failure cases of the approach and "MHCF

counter measures to failure cases" discusses the counter measures. This is followed by Novelty of MHCF, followed by the chapter summary.

4.1 Proposed Approach: Formal Definition

Given a set of papers $P = \{p_1, p_2, ...p_m\}$, the system needs to group all papers authored by the same author within a cluster. The result is a set of clusters C $= \{c_1, c_2, c_3, c_4, \ldots, c_n\}$, (where $n \leq m$), and each cluster c_g represents a distinct author ($c_g \in C$, where $1 \leq g \leq n$).

Two papers p_i and $p_j \in P$, (where, $1 \leq i < m$, $i < j \leq m$), should belong to the same cluster (c_g), if $F_{similarity}(p_i, p_j) \rightarrow 1$ (normalized value), then $(p_i, p_j) \in c_g$. $F_{similarity}$ is a function that checks the similarity of any two papers given a feature or combination of features. Therefore, given a set of ambiguous papers P, it is intended to: group all papers belonging to a distinct author within a cluster to achieve an increased overall pairwise F1 [1].

4.2 Multilayer Heuristic Based Clustering Framework (MHCF)

4.2.1 Rationale of using Heuristics Based Unsupervised Learning for MHCF

MHCF is majorly inspired by the heuristic based clustering strategies commonly used by researchers in AND. Before discussing the approach, the rationale behind using this approach is that different approaches have been proposed over time to better solve this problem, which mainly uses supervised, unsupervised, or graph-based learning. It is observed through the literature review that among supervised

[1] Appendix D covers an example scenario which addresses the effect of precision, recall and F1 scores if a paper is authored by multiple authors

and graph based models, techniques which are based on unsupervised learnings, are intuitively more suitable to resolve AND, an observation also pointed out by Pooja et al and Z. Zhao et al [46, 71] in their recent studies.

Unsupervised approaches have an edge over supervised learning based techniques because they can easily gather samples with similar characteristics together without knowing the number of classes in advance. Similarly, these approaches do not need labeled training sets to perform the said task. Moreover, it is more likely that unsupervised models can disambiguate non-active[2] researchers as opposed to supervised learning models, without needing large training sets. Similarly, in comparison to graph based models, unsupervised approaches usually do not rely on a specific feature (often co-authors) to construct the heterogeneous networks.

Based on these grounds, though, the proposed approach is inspired by classic hierarchical agglomerative clustering (HAC), however, it differs from it and other heuristics-based approaches as, 1) instead of merely relying on proximity matrix calculations to merge two groups, MHCF uses set of heuristics corresponding to different features (one at a time till no further merging is possible against a single feature) at the instance level to intelligently merge two instances or groups. 2) As opposed to existing AND approaches, MHCF is simple, flexible (in terms of use of any available feature yet keeping in view its usefulness towards attaining better F1 scores) and considers the contextual meaning of words, rather than keywords existence in certain features.

4.2.2 Components of the Framework and their Working

This section illustrates the components of MHCF, along with their working, heuristics, and proposed algorithms. MHCF is majorly composed of three layers. The first layer is the initial layer which organizes the papers in blocks. The next is the structure-aware layer, i.e. layer 2, which uses the structure-aware features to cluster the data.

[2]Authors whose number of publications is less

Layer 3 is the global feature layer which is responsible for merging clusters using global features. (It is worth mentioning here that at each layer, multiple features are considered by MHCF to enhance AND, however, their use is subjective to their availability in the dataset that is being used for this process).

To elaborate and discuss each layer in detail, consider the following subsections as described below.

4.2.2.1 Layer 1

The **First layer** is considered as the preliminary layer of MHCF, which groups the input papers into a set of ambiguous blocks ($B = \{b_1, b_2, \ldots, b_z\}$, where $z \geq 1$). All papers with the same first name initial and last name of authors are grouped to form an ambiguous block.

Each paper within a block b_t ($b_t \in B$, such that $1 \leq t \leq z$ and z = number of papers in that block) is treated as a separate cluster. *(Clusters and groups will be used interchangeably in the rest of this chapter)*

4.2.2.2 Layer 2

The second layer is the **structure layer**, in which MHCF tries to merge two clusters based on the structure-aware features in an incremental fashion. The structure aware features are discussed in Chapter 3, section 3.3.3.1, whereas the sequence with which they are used in the layer is based on the proposed feature ranking in Chapter 3, i.e. author email, author affiliation, co-authors, followed by features which come under layer 3, which are followed by using paper venue feature.

Layer 2 (structure layer) is responsible for merging individual groups by gauging them using a set of heuristics. Use of these features can be skipped in case of their unavailability in the test dataset, except co-authors feature, which is the very basic feature of this layer. The main reason for this is that except single-authored

publications, this feature value is likely to be present in almost all publications, as compared to other structure aware features. Heuristics with respect to each feature are illustrated in the sub-sections as under:

Clustering Based on Author Emails

To start off, MHCF merges two groups G_i and G_j ($1 \leq i \leq n-1, i+1 \leq j \leq n$, where n = number of clusters/groups) if and only if they share one or more authors' emails based on the cosine similarity score greater than the threshold i.e., 0.8.

Algorithm 1 achieves the structure-aware clustering using author's email feature following the heuristics as discussed above.

Algorithm 1 authorEmailBasedClustering

Input: G

Global: G

Output: G

```
1: while (no more group mergence is possible) do
2:     for (i in range (0: len(G)-1)) do
3:         for (for j in range (i+1: len(G))) do
4:             if (emails of two groups G_i and G_j are similar) then
5:                 merge G_i and G_j groups
6:             end if
7:         end for
8:     end for
9: end while
```

Clustering Based on Author Affiliations

After making clusters using author email, MHCF tries to merge two groups G_i and G_j, if and only if the two groups share one or more author affiliations, based on cosine similarity score greater than the threshold value 0.8.

Algorithm 2 is related to author affiliation-based clustering.

Algorithm 2 authorAffiliationBasedClustering

Input: G

Global: G

Output: G

```
1: while (no more group mergence is possible) do
2:     for (i in range (0: len(G)-1)) do
3:         for (j in range (i+1: len(G))) do
4:             if (affiliations of two groups G_i and G_j are similar) then
5:                 merge G_i and G_j groups
6:             end if
7:         end for
8:     end for
9: end while
```

Clustering Based on Co-authors Names

MHCF uses its own author name comparison algorithm to find common co-authors in the groups. Two groups are merged if and only if they have one or more than one overlapping co-authors among them. Initially, co-authors name are matched based on exact matches without considering misspelling and name phoenix. In case of no match, MHCF checks the name fragments of co-authors. For this, co-authors with equal name fragments[3] (i.e. has same first name or first three characters of the first name, and same last name), is considered same, if and only if they also share same middle names (either full middle name or first character of the middle name). In case of no common co-authors, MHCF checks and considers two authors same, if they have unequal but at least three name fragments, they

[3]an author will have 3 name fragments if s/he has first name, middle name, and last name

have same first name or first three characters of the first name or first character of the first name, they have same middle name (full name or first character) and have full last name. Similarly, authors with same first and last names are considered same if and only if their remaining name fragments first characters are also same. (All single authored papers and papers with no co-authors are rejected at this phase, which are catered using other features).

Algorithm 3 is related to structure-aware clustering using coauthors name feature. Two groups are merged if they have one or more than one common coauthors (based on exact match). If there exists no such match, merge the clusters based on their name fragments matching.

Algorithm 3 coauthorsBasedClustering

Input: G

Global: G

Output: G

```
1: while (no more group mergence is possible) do
2:     for (i in range (0: len(G)-1)) do
3:         for (j in range (i+1: len(G))) do
4:             if (coauthors of two groups G_i and G_j are same) then
5:                 merge G_i and G_j groups
6:             else if (fragments of coauthors of G_i and G_j are same) then
7:                 merge G_i and G_j groups
8:             end if
9:         end for
10:    end for
11: end while
```

Clustering Based on Paper Venue

MHCF merges two groups/clusters G_i and G_j if and only if they share one or more paper venues based on cosine similarity score greater than the threshold and have

at least one common coauthor. The initial threshold value is set to 0.8 which is decreased to 0.5 if the two groups share common co-authors.

Algorithm 4 illustrates the structure aware clustering using the paper venue feature.

Algorithm 4 paperVenueBasedClustering

Input: G, threshold = 0.8

Global: G

Output: G

```
1:  while (no more group mergence is possible) do
2:      for (i in range (0: len(G)-1)) do
3:          for (for j in range (i+1: len(G))) do
4:              if (same coauthors count in two groups G_i and G_j ≥ 1) then
5:                  threshold ← 0.5
6:              end if
7:              if (venue of group G_i and G_j are similar) then
8:                  merge G_i and G_j
9:              end if
10:         end for
11:     end for
12: end while
```

4.2.2.3 Layer 3

The third layer is responsible of using the proposed Research2Vec model to extract global features from the paper title, paper abstract and paper keywords, in the **global layer**, one by one. The details regarding the proposed model Research2Vec and global features are given in Chapter 3, section 3.3.3.1. Whereas the sequence of use of features by MHCF in the global layer is based on the F1 scores discussed in Chapter 3, i.e. paper title, paper abstract, and paper keywords.

Clustering Based on Paper Titles, Abstracts, and Keywords

Each cluster will have one combined paper title giving a combined representation of all the papers titles within that cluster. Similar global features are extracted against two clusters using Research2vec model. Research2vec takes tokenized title words giving a feature vector against each word, which are averaged together to yield one averaged vector against each input string. Two clusters are merged if the averaged vectors' cosine distance is within a given threshold and the groups have at least one common coauthor. Initially, the threshold is set to 0.85 which is decreased to 0.5 if the groups share common co-authors. A similar procedure is adopted for paper keywords and paper abstracts.

Algorithm 5 illustrates the global features based clustering using the paper title feature.

Algorithm 5 paperTitleBasedClustering

Input: G, threshold = 0.85

Global: G

Output: G

```
1: while (no more group mergence is possible) do
2:     for (i in range (0: len(G)-1)) do
3:         for (j in range (i+1: len(G))) do
4:             if (same coauthors count in two groups G_i and G_j ≥ 1) then
5:                 threshold ← 0.5
6:             end if
7:             if (title similarity of two groups G_i and G_j ≥ threshold)) then
8:                 merge G_i and G_j groups
9:             end if
10:        end for
11:    end for
12: end while
```

Two groups are merged if their word embedding vectors are similar. The default threshold value used to compare two vectors is initially set to 0.85, which can be relaxed to 0.5 if the two groups share one or more common coauthors between them.

Algorithm 6 is related to global features-based clustering using the paper abstract feature. Two groups are merged if their abstract word embedding vectors are similar. Finding the word embeddings and initial threshold values versus threshold relaxation criteria is similar to the paper title feature.

Algorithm 6 paperAbstractBasedClustering

Input: G, threshold = 0.85

Global: G

Output: G

```
1: while (no more group mergence is possible) do
2:     for (i in range (0: len(G)-1)) do
3:         for (j in range (i+1: len(G))) do
4:             if (same coauthors count in two groups G_i and G_j ≥ 1) then
5:                 threshold ← 0.5
6:             end if
7:             if (abstract similarity of two groups G_i and G_j ≥ threshold)) then
8:                 merge G_i and G_j groups
9:             end if
10:        end for
11:    end for
12: end while
```

Algorithm 7 is related to the paper keyword-based clustering.

Two groups are merged if their keyword word embedding vectors are similar. Finding the word embeddings and initial threshold values vs threshold relaxation criteria is similar to the paper title feature.

Algorithm 7 paperKeywordBasedClustering

Input: G, threshold = 0.85

Global: G

Output: G

```
1: while (no more group mergence is possible) do
2:     for (i in range (0: len(G)-1)) do
3:         for (j in range (i+1: len(G))) do
4:             if (same coauthors count in two groups G_i and G_j ≥ 1) then
5:                 threshold ← 0.5
6:             end if
7:             if (keyword similarity of two groups G_i and G_j ≥ threshold)) then
8:                 merge G_i and G_j groups
9:             end if
10:        end for
11:    end for
12: end while
```

4.2.3 Putting MHCF into Work

The initial step of MHCF before making ambiguous blocks is pre-processing of the data. To pre-process the data, first, the stop words are removed from the textual features. This is followed by changing the alphabetical characters to lower-case and by changing the character encoding scheme to ASCII format. The encoding scheme is used to make the data consistent. Next, all missing values are replaced by "none" and any word including: {'issues', 'international', 'proceedings', 'proceeding', 'journal', 'conference', 'conferences', 'workshop', 'workshops', 'proc.', 'symposium' etc}, are removed from the paper publishing venue feature.

MHCF uses Algorithm 8, as the main entry point to the proposed system. The algorithm takes ambiguous publications to disambiguate and reference clusters to

check the output i.e., systems-generated clusters' accuracy. This algorithm starts with the pre-processing of the publications. This is followed by making ambiguous blocks based on the author's first name initial and last name. Next, for each block, make groups such that each group has a single paper in it. After this, the groups are merged based on structure-aware and global-aware features, incrementally. All of this is followed by result calculations, in which the results are calculated and stored per block, which is finally used to calculate the overall results of MHCF. Figure 4.1 gives an overview of MHCF.

Algorithm 8 MHCF

Input: Set of ambiguous publications P, Reference clusters R

Global: G, B, R

\\G is a set of clusters, B is a set of ambiguous blocks, R is a set of block wise results

Output: Set of system generated global clusters G

1: $P \leftarrow$ pre-process publications(P)
2: $B \leftarrow$ make ambiguous blocks(P)
3: **for each** B **do**
4: $G \leftarrow$ make groups
\\make groups such that each group has one paper in it.
\\use the features based on their availability in the dataset.
5: $G \leftarrow$ authorEmailBasedClustering (G)
6: $G \leftarrow$ authorAffiliationBasedClustering (G)
7: $G \leftarrow$ coauthorsBasedClustering (G)
8: $G \leftarrow$ paperTitleBasedClustering (G)
9: $G \leftarrow$ paperAbstractBasedClustering (G)
10: $G \leftarrow$ paperKeywordBasedClustering (G)
11: $G \leftarrow$ paperVenueBasedClustering (G)
12: R $\leftarrow$ calculate and save results (G)
13: **end for**
14: over_all_pP, pR, pF1, ACP, AAP, K $\leftarrow$ calculate overall results (R)

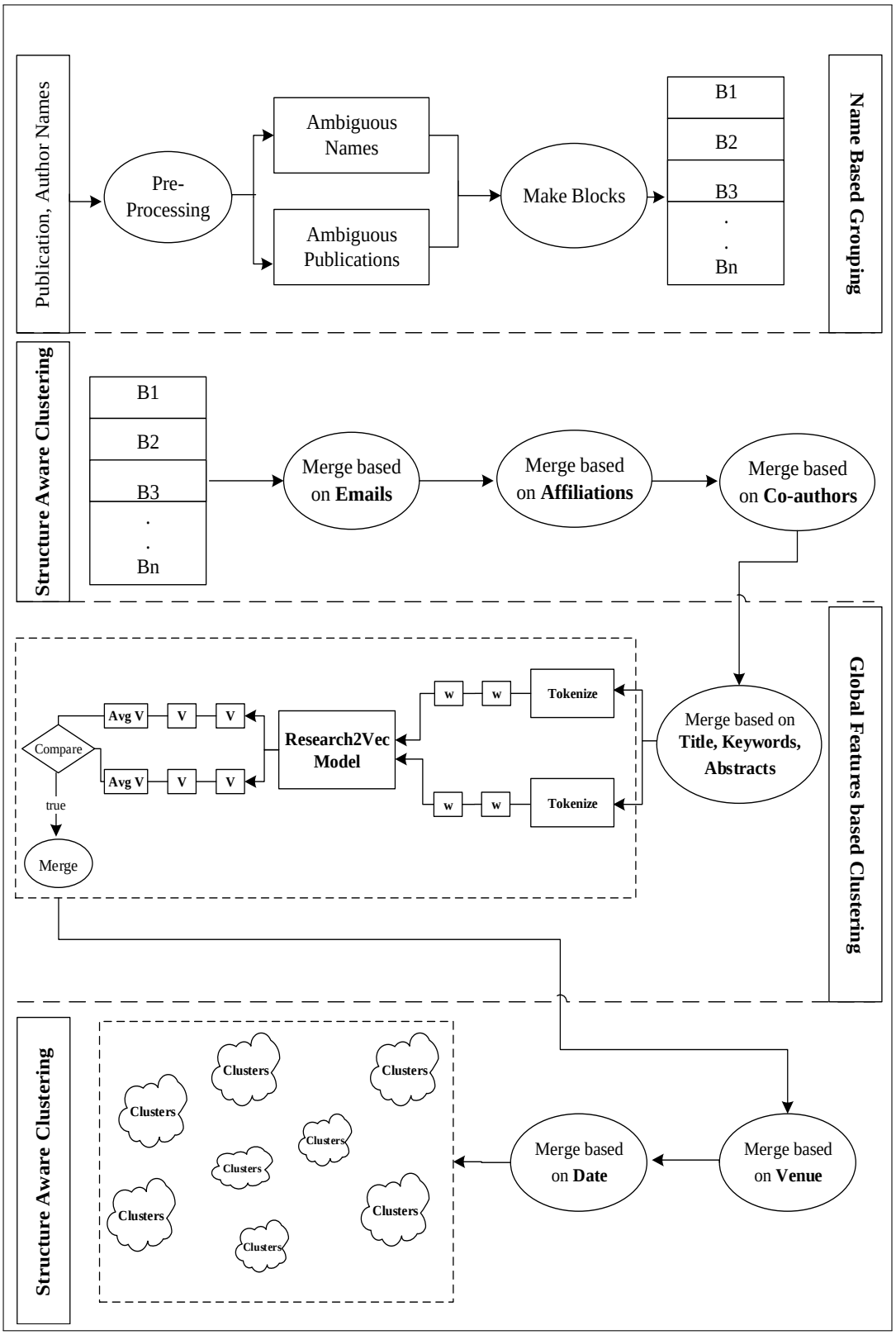

FIGURE 4.1: MHCF Workflow

In MHCF, the cluster mergence based on a single feature either on the structure

layer or the global layer will also merge the rest of the feature's values without assessing them. For instance, two clusters that are merged based on common co-authors features in the structure layer will also result in blindly merging the rest of the feature values.

After each merge, the system repeats the process until no further merging is possible with respect to the feature in use. The system then takes the next feature and repeats the process. Therefore, after making ambiguous author name blocks, the clusters are merged based on structure-aware features, including author email, author affiliation, and co-authors. This is followed by merging the resultant groups based on the paper title, paper abstract, and paper keywords. Lastly, the resultant groups are merged based on venue and publishing date.

As far as the **threshold value selection criteria** is concerned, feature-wise threshold selection is performed by gauging a threshold sensitivity analysis. An experiment was performed in which varied threshold values are used, starting from 0.1 to 1.0 against each feature. This study performed 10 executions of the method and recorded the results. Threshold values that maximize the output are finally selected. Threshold value for the paper title, abstract, and keywords is 0.85. For author's email, affiliation, and paper venue features, the threshold value is 0.8.

The next section discusses the experimental setup and MHCF results.

4.3 Experimental Setup and Results

4.3.1 Datasets

The proposed framework is evaluated on ArnetMiner[4] and BDBComp[5]. Arnetminer dataset is created by Wang et al, [25], and contains authorship records that are extracted from the data which is collected within Arnetminer[6] system.

[4]http://arnetminer.org/disambiguation

[5]http://lbd.dcc.ufmg.br/bdbcomp

[6]Table A.19 show the details regarding ambiguous author groups per the number of references, and the number of distinct authors per group in the BDBComp

Authorship records in this collection are associated with 109 ambiguous authors. Whereas, the collection of citations extracted from BDBComp[7] sums up to 361 citations which are associated with 205 distinct authors with eight author names in short format.

4.3.2 Baselines

For the baselines, two variations of MHCF are considered i.e., MHCF-G and MHCF-GL. MHCF-G uses a pre-trained word2vec model using Wikipedia corpus with vocabulary size 2000000. MHCF-GL uses GloVe word embeddings using Stanford pre-trained model using Wikipedia 2014 corpus [93]. The two variations (MHCF-G and MHCF-GL) use pre-trained word embedding models to extract global features from paper titles, keeping everything else the same as in MHCF. Comparison of MHCF with its own variations will give the insights to the worth of Research2Vec embedding model as compared to the existing ones. In addition to MHCF-G and MHCF-GL, two hybrid models are also selected, SAND1 [21], SAND2 [22], which use self-trained data to identify the output class, where the trained data is established by clustering set of papers based on overlapping co-authors. Similarly, HHC [20], which is an unsupervised heuristics based hierarchical clustering technique similar to MHCF is selected, along with a novel multiple layers name disambiguation framework [18] (which will be referred to as MDC in the rest of the chapter). MDC adopts a dynamic clustering mechanism to minimize clustering errors using multiple features. In MDC, co-authors-based merging is done using Erdos number theory whereas paper titles-based merging is done using Gensim based topic modeling. MHCF is compared to these techniques as their approach is quite similar, with slight variations. Another technique (GFAD) which is proposed by Shin et al [17] is selected for comparison with MHCF. It is a graph based technique that solves the under-discussed problem by splitting an author vertex involved in multiple cycles of co-authorship, and, by merging

[7]Table A.18 show the details regarding ambiguous author groups per the number of references, and the number of distinct authors per group in the BDBComp

multiple author vertices having similar names if they are connected to a common vertex.

MHCF is also compared with two hybrid graph-based approaches proposed by P.Km etal and Pooja et al., [8, 19] respectively. In the first approach (ATGEP [19]), two different components of the author-author graph are merged, if at least one document from each of the two components appears in the publication profile of the author which is created using external web sources. Whereas, in the second approach (It will be referred to as - ESMD [8], in this chapter), the authors use unsupervised learning with graph autoencoders to embed different feature values i.e. co-authors, paper title, abstract, venues, references, and affiliation. The rationale behind the selection of these techniques for comparison with MHCF is to justify MHCF's strength with different approaches other than clustering. The mentioned techniques range from simple graphs to more complex graph-based techniques.

To evaluate MHCF with SAND1, SAND2, HHC, MDC, GFAD, ATGEP, and ESMD, their reported results are used, as all the settings are kept the same to get MHCF results.

4.3.3 Evaluation Metrics

To validate the clusters, pairwise precision, pairwise recall, and pairwise F1 scores are utilized, as recommended by A. Elke et al. [94].

To measure the efficacy of the solution artifacts this study uses a variety of evaluation metrics that are commonly used in AND perspective, to measure the AND techniques results [30], as, they give an insight into the cluster's purity, and their cohesion factor.

Ideally, the AND techniques should combine the publications authored by the same author into one cluster and segregate publications authored by others into their respective clusters [30].

4.3.3.1 Pairwise Precision (pP)

pP is calculated by computing the authorship record pairs in a predicted cluster that are correctly associated with the same author as compared to the number of authorship record pairs in a predicted cluster not corresponding to the same author [30]. It is computed using the formula given as follows:

$$\mathrm{pP} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{Q}} \sum_{\mathrm{j}=1}^{\mathrm{R}} \mathrm{C}(\mathrm{n_{ij}}, 2)}{\sum_{\mathrm{i}=1}^{\mathrm{Q}} \mathrm{C}(\mathrm{n_i}, 2)} \tag{4.1}$$

Where: Q is the predicted clusters, R is the reference clusters for this ambiguous group, n_{ij} is the total number of authorship records in the predicted cluster i that are also in the reference cluster j, and n_i is the total number of authorship records in the predicted cluster i.

C(n, r) denotes the number of combinations of r elements from n elements as given:

$$\mathrm{C}(\mathrm{n}, \mathrm{r}) = \frac{\mathrm{n!}}{\mathrm{r!.(n-r)!}}, \mathrm{n} \geq \mathrm{r} \tag{4.2}$$

4.3.3.2 Pairwise Recall (pR)

pR is calculated by computing the number of authorship record pairs associated with the same author that are not in the same predicted cluster [30], and is given as:

$$\mathrm{pR} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{Q}} \sum_{\mathrm{j}=1}^{\mathrm{R}} \mathrm{C}(\mathrm{n_{ij}}, 2)}{\sum_{\mathrm{j}=1}^{\mathrm{Q}} \mathrm{C}(\mathrm{n_j}, 2)} \tag{4.3}$$

4.3.3.3 Pairwise F1 (pF1)

pF1 is the F1 metric calculated using pP and pR [30] following the equation:

$$\mathrm{pF1} = \frac{2.\mathrm{pP.pR}}{\mathrm{pP} + \mathrm{pR}} \tag{4.4}$$

4.3.3.4 ACP Metric

ACP measures the quality of clustering by calculating the average purity of the clusters formed. Purity refers to the percentage of authors in a cluster that belongs to the same ground-truth identity [30]. If the predicted clusters are pure, the corresponding ACP value will be 1. ACP is given by the following equation:

$$\mathrm{ACP} = \frac{1}{\mathrm{N}} \sum_{(\mathrm{i}=1)}^{\mathrm{Q}} \sum_{(\mathrm{j}=1)}^{\mathrm{R}} \frac{\mathrm{n}_{\mathrm{ij}}^2}{\mathrm{n}_{\mathrm{i}}} \tag{4.5}$$

Where N is the total number of academic authorship records in the ambiguous group, Q is the predicted clusters, R is the reference clusters for this ambiguous group, n_{ij} is the total number of authorship records in the predicted cluster i that are also in the reference cluster j, and n_i is the total number of authorship records in the predicted cluster i.

4.3.3.5 AAP Metric

AAP measures the quality of disambiguation by calculating the average precision of a system over all authors in a test dataset. Precision refers to the percentage of authors that are correctly disambiguated to their ground-truth identity. AAP provides insight into the overall accuracy of a disambiguation system, regardless of clustering [30]. AAP is given by equation as follows:

$$\mathrm{AAP} = \frac{1}{\mathrm{N}} \sum_{(\mathrm{j}=1)}^{\mathrm{R}} \sum_{(\mathrm{j}=1)}^{\mathrm{Q}} \frac{\mathrm{n}_{\mathrm{ij}}^2}{\mathrm{n}_{\mathrm{j}}} \tag{4.6}$$

Where n_j is the total number of authorship records in the reference cluster j.

4.3.3.6 K Metric

K metric is the geometric mean between ACP and AAP values. It evaluates the purity and fragmentation of the predicted clusters identified by a specific

disambiguation method [30]. The K metric is given by the equation as:

$$K = \sqrt{ACP.AAP} \tag{4.7}$$

4.3.3.7 Cluster Precision (CP)

CP is the fraction of correct clusters as compared to the incorrect ones. A cluster is correct if it has all the authorship records of an author and none from another author [30]. It is calculated as:

$$CP = \frac{a}{a + c} \tag{4.8}$$

Where a is the number of correct clusters (a correct cluster should have all the authorship records of an author and only those, i.e., none from another author, otherwise it is incorrect). Whereas c is the number of incorrect clusters.

4.3.3.8 Cluster Recall (CR)

CR is the fraction of correctly predicted clusters compared to the reference clusters [30]. The calculations are done using:

$$CR = \frac{a}{a + b} \tag{4.9}$$

Where b is the number of clusters that should be created but were not.

4.3.3.9 Cluster F1 (CF1)

CF1 is the harmonic mean of CP and CR, where CP is the fraction of correct clusters as compared to incorrect clusters. A cluster is correct if it has all the authorship records of an author and none from another author, whereas CR is the

fraction of correctly retrieved clusters compared to the reference clusters [30]. It is computed as:

$$\mathrm{CF1} = 2.\frac{\mathrm{CP.CR}}{\mathrm{CP} + \mathrm{CR}} \tag{4.10}$$

4.3.3.10 RCS

RCS is given by dividing the number of predicted clusters by the reference ones. This serves to evaluate how close is the measure to the ideal number of clusters to be generated [30].

4.4 Results

This section covers the results of MHCF in comparison to the baseline approaches. The detailed results of MHCF in comparison to all the baseline techniques are reported individually in Appendix A. Whereas Table 4.1 lists the pF1 scores of MHCF along with all the baselines.

TABLE 4.1: MHCF pF1 Results Comparison using Arnetminer and BDBComp.

	BDBComp	**Arnetminer**
Technique	**pF1**	**pF1**
MHCF	**0.86**	**0.88**
MHCF-G	0.83	0.84
MHCF-GL	0.85	0.8
SAND1	0.68	-

SAND2	0.75	-
HHC	0.65	-
MHCF(40 names, 2 features)	-	**0.81**
GFAD-OR	-	0.75
GFAD-AD	-	0.75
MHCF(11 names, 2 features)	-	**0.73**
MDC(11 names, 2 features)	-	0.65
MHCF(11 names, 3 features)	-	**0.77**
MDC(11 names, 3 features)	-	0.75
MHCF(15 names)	-	**0.95**
ESMD	-	0.91
ATGEP	-	0.71

Overall, the results show that MHCF with optimal threshold values and no prior knowledge of k (number of clusters) give better pF1 performance than SAND1[8] [21], SAND2[9] [22], HHC [20], GFAD[10] [17] MDC[11] [18], ATGEP[12] [19] and ESMD [8], owing to the capability to incorporate contextual information, rather than relying on the presence of same keywords in the text. Additionally, the heuristics and prioritized use of powerful discriminating features help MHCF to achieve better pP results by minimizing the inclusion of false positives in the early steps

[8]Table A.21
[9]Table A.21
[10]Table A.23
[11]Table A.24 reports MHCF detailed results using two features(co-authors and paper title) and three features(o-authors, author affiliation, paper title), against 11 ambiguous author names, in comparison to the MDC technique.
[12]The results in Table A.25 (Appendix A) show that MHCF also achieves better pP results as compared to ESMD[13] and ATGEP

and avoiding the replication of errors later. This behavior achieves better overall results without compromising precision.

4.5 Analysis of the MHCF Results

In the efforts to explain the performance of MHCF discussed in the previous sections, first, the failure cases of MHCF are discussed, along with the intuition of why it occurred. In addition, after discussing the drawbacks and reasons for failure cases, another experiment is conducted to see whether the counter arguments which are given against each failure case are correct or not. For this purpose, MHCF is evaluated on the proposed dataset CustAND (chapter 5) which is publicly available online.

4.5.1 Low Precision (Arnetminer perspective)

To analyze low precision by MHCF few low pP achieving names are selected from 109 author names[14] of Arnetminer dataset. The first doubt about the low pF1 achieving name blocks is that either these blocks have some missing feature values or the use of some features pollutes the cluster purity. For this purpose, few statistics corresponding to the selected names are given in Table 4.2. The statistics support the first doubt and show that the author affiliation feature has the highest missing values which is followed by paper venues that make it difficult to precisely distinguish and correctly cluster the data.

To extend the analysis and find evidence against the second intuition i.e. "use of some features pollutes the cluster purity", features contributions are analyzed within the selected blocks. For this, the features are eliminated one by one in the reverse order from case_1 denoted as case_2, case_3, and case_4 (to see low precision's relation with different features) as shown in Table 4.3. This means that case_2 involves co-authors name, author affiliation, and paper title features,

[14]Table A.26 of Appendix A

case_3 involves co-authors name and author affiliation, and case_4 includes only co-authors feature. The results under case_2 (column) show that the pP increased by a significant number just by excluding the paper publishing venue from "Bo Liu" in which only 8% venues were missing, whereas 73% records have missing affiliations and have no single authored publication records. Similarly, the exclusion of paper title feature increased the pP to 96%, which goes up to 100% by using only co-authors name feature as shown under case_4. A similar phenomenon is witnessed in other selected name blocks. This concludes that in these groups paper venue feature seems to be the most polluting feature compromising the overall pP, which ultimately lowers the overall pF1 score.

TABLE 4.2: MHCF Lowest pP Achieving Author Blocks.

Ambiguous Name Blocks	**Total Records**	**Missing affiliations**	**Missing Venue**	**Single authored publications**
Bo Liu	124	73%	8%	0%
Bin Li	181	29%	6%	1%
Feng Liu	149	48%	5%	3%
Gang Chen	178	36%	5%	3%
Jing Zhang	231	37%	10%	3%
Ke Chen	107	56%	9%	13%
Lei Wang	308	45%	9%	3%
Yang Wang	195	48%	4%	7%
Yu Zhang	235	40%	5%	6%
Paul Brown	27	81%	19%	26%
Bin Zhu	46	72%	7%	2%

To further analyze the low precision problem in the selected name blocks, and to determine whether the performance of MHCF as a framework is at fault or if the dataset is inadequate, two failure cases are discussed.

Table 4.3: MHCF Lowest pP Achieving Author Blocks.

Authors	**Case_1 (4 features)**			**Case_2 (3 features)**			**Case_3 (2 features)**			**Case_4 (1 features)**		
	pP	**pR**	**pF1**	**pP**	**pR**	**pF1**	**pP**	**pR**	**pF1**	**pP**	**pR**	**pF1**
Bo Liu	0.28	0.99	0.43	0.76	0.81	0.78	0.96	0.81	0.88	1	0.56	0.71
Bin Li	0.54	0.93	0.68	0.95	0.93	0.94	0.99	0.93	0.96	1	0.79	0.88
Feng Liu	0.58	0.45	0.5	1	0.42	0.59	1	0.42	0.59	1	0.39	0.56
Gang Chen	0.49	0.77	0.6	0.77	0.48	0.59	0.77	0.48	0.59	1	0.48	0.65
Jing Zhang	0.09	0.78	0.17	0.86	0.74	0.79	0.86	0.74	0.79	0.98	0.5	0.66
Ke Chen	0.59	0.6	0.59	1	0.37	0.54	1	0.37	0.54	1	0.29	0.46
Lei Wang	0.07	0.93	0.13	0.58	0.86	0.69	0.68	0.86	0.76	0.87	0.71	0.78
Yang Wang	0.29	0.54	0.38	0.99	0.44	0.61	0.99	0.44	0.61	0.98	0.3	0.46
Yu Zhang	0.48	0.65	0.55	0.96	0.6	0.74	0.96	0.53	0.68	1	0.55	0.71
Paul Brown	0.51	0.76	0.61	1	0.55	0.71	1	0.55	0.71	1	0.66	0.8
Bin Zhu	0.57	0.74	0.64	1	0.74	0.85	1	0.74	0.85	1	0.51	0.67

4.5.1.1 Low Precision due to Shared Venues Among More Than One Distinct Author (Failure Case 1)

One possibility behind low pP achieving ambiguous name blocks, after using the paper venue feature is that distinct authors within these groups either share their publication venues with each other or the same publication venue acronyms point to different venues. This is confirmed by seeing the word graph of some of the blocks including "Gang Chen", "Paul Brown" and "Bin Zhu" as shown in Figure 4.2.

For instance, "CSCWD" venue acronym is shared by more than one distinct author in "Gang Chen" block, "SIGMOD" and "VLDB" is shared between different authors in "Paul Brown" block, and "ICIP" is shared between distinct authors in "Bin Zhu" ambiguous block, etc. Though other blocks may suffer from this problem too, it is quite possible that the number of distinct authors sharing common venues in other blocks is low.

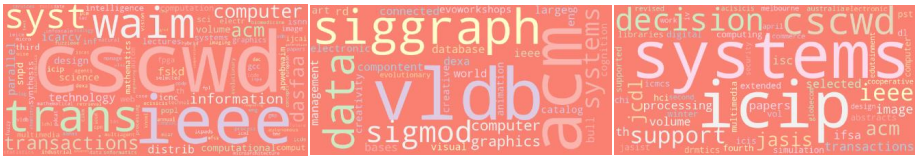

FIGURE 4.2: Word Graph of "Gang Chen", "Paul Brown" and "Bin Zhu".

4.5.2 Low Recalls (Arnetminer Perspective)

This section discusses the low recall cases in the ambiguous blocks.

4.5.2.1 Missing Feature Values and In-availability of Features (Failure Case 2)

This case occurs when the features that give high pF1 scores (identified in Chapter 3) are either not available or have sparse values. For example, 1) author email

feature is absent. 2) sparse availability of author affiliation values in the dataset. 3) cases of single authored papers, where their instances have sparse feature values too 4) no similar global features are encountered during the cluster mergence phase.

All of this ultimately leads to low recalls in MHCF. Though, using paper venue feature increases the overall recall of the group but this feature has the capability to pollute the cluster precision(Chapter 3).

4.5.3 Low Precision and Recall (BDBComp perspective)

BDBComp is one of the toughest datasets among all the available ones as it has only 3.47 publications which are associated per distinct authors. Also, paper titles and paper venues are mostly based on non-English characters and give a low feature coverage i.e., provide only three features per record to work with. Table 4.4 shows the overall statistics of the co-author's names per record per ambiguous author name in BDBComp. It can be seen that on average 71% of the total co-author names have two name fragments which have only first name initial and full last names. Such name variations are ambiguous and need supplementary evidence to discriminate them, which is either absent or is limited in the collection. For example, author email, author affiliation features are absent. Whereas, author full names, though is an ambiguous feature however, can contribute to attain better result in case if the author name is not shared with others. Similarly, co-authors names have a very low percentage of middle name availability, which itself is not a good feature but can sometimes facilitate the technique in AND.

TABLE 4.4: Co-authors Names Statistics in BDBComp.

Names	**Total co-authors**	**Co-authors with first name initial, last name**	**Co-authors with first, middle and last name**
a oliveira	174	72%	3%
a silva	232	72%	2%
f silva	89	70%	1%

j oliveira	159	69%	1%
j silva	116	70%	2%
j souza	125	72%	0%
l silva	112	71%	0%
m silva	78	73%	4%
r santos	86	77%	2%
r silva	91	69%	2%

Majorly MHCF low performance on BDBComp can be complemented to the fact that Research2vec does not contribute at all in cluster refinement after applying coauthor-based merging, as the model is trained on English research articles only. Since majority of the titles in the BDBComp collection have non-English titles, MHCF is unable to find any word embeddings against them, thus limiting MHCF overall performance.

4.6 MHCF Counter Measures to Failure Cases

Failure case 1 is not related to MHCF incapability, so it is not considered. Failure case 2 though is not directly related to MHCF incapability to perform the said task, so it can be re-checked using some other data collection, which gives better feature coverage and has complete values against the instances. To do so, MHCF is evaluated on another dataset with complete information against the same features, which are used while considering Arnetminer collection i.e., paper title, co-authors, author affiliation, and paper venue. MHCF is also evaluated using a different feature combination, which is identified to produce better results as given in Chapter 3. The next section shares its details and results.

4.6.1 Performance Evaluation of MHCF with 'CustAND' for Failure Case 2.

For this purpose, CustAND dataset is used [15], where its details can be seen in Chapter 5. Table 4.5, shows the overall MHCF pF1 performance using CustAND with complete feature values. The results show that complete feature values can improve MHCF performance. Similarly, MHCF performance using different feature combinations (author emails and author affiliations) shows that overall better pF1 can be achieved with less but better result producing features combination. Therefore, this experiment supports the intuitions given against failure case 3 i.e., a flexible AND framework can perform better if given complete and relevant information rather than incomplete or such features set, which inversely affects the results.

Table 4.5: Overall MHCF Results using CustAND Data Collection.

pP	**pR**	**pF1**	**ACP**	**AAP**	**K**	**Features Used**
94.60%	92.50%	93.50%	95.80%	87%	91.24%	Co-authors, author affiliations, paper title, paper venue
100%	96%	**98%**	99%	97%	98%	Author email, author affiliation

Similarly, the use of features that have less probability to be shared among distinct authors can also help to achieve better results as compared to the ones which can be shared among multiple authors; therefore, this experiment also encounters failure case 1.

4.7 Novelty of MHCF

MHCF algorithm is a novel approach which addresses the author name ambiguity problem, with a significant improvement in precision, recall and F1 scores as

compared to other techniques. The novelty of MHCF is articulated through the following key contributions:

1. The MHCF algorithm introduces a novel approach to Author Name Disambiguation (AND) by incrementally employing ranked features, complemented by intelligently designed rules for cluster formation. This methodology stands out for its capability to significantly eliminate false positives during the initial merging stages, giving a more accurate disambiguation process compared to other techniques. Moreover, MHCF adopts a layered merging strategy that leverages ranked features and combinations, aiming to carefully increase the recall score, while limiting the false positives with each iterative merge using progressively lower ranked features (descending order of their rank).

2. MHCF utilizes the proposed novel Research2Vec embedding model, which is trained on the arXiv dataset encompassing academic papers from diverse domains. This embedding model is publicly available for research purposes, offering vectors that exhibit significantly enhanced semantic relevance compared to pre-trained Word2Vec model on Wikipedia articles. The Research2Vec-based semantically coherent vectors play a vital role in reducing false positives when used with features like paper titles, abstracts, and keywords by MHCF. This enhancement is empirically validated through comparative evaluations with alternative models, including MHCF-G (utilizing pre-trained Word2Vec on Wikipedia articles) and MHCF-GL (employing GloVe embeddings with Stanford's pre-trained model on the Wikipedia 2014 corpus).

3. Comprehensive experimentation's demonstrate that MHCF algorithm achieves significant improvements in precision, recall, and F1 scores when compared to other existing AND techniques.

4.8 Chapter Summary

This chapter focuses on discussing the proposed Multilayer heuristics-based clustering framework to improve the authorship results (in the form of better F1 scores). MHCF is inspired by classic hierarchical agglomerative clustering, where it uses heuristics per feature at the instance level to merge two groups, instead of merely relying on proximity matrix calculations as in case of traditional HAC.

MHCF uses features in accordance with their power to identify distinct authors along with the contextual as well as non-contextual features using the proposed structure-aware and global feature layer to group papers. The use of the proposed Research2Vec model to extract global features by MHCF helps to achieve reasonable recall with fewer false positives as compared to similar existing techniques. Whereas, ranked use of structure-aware features help in achieving pure clusters. MHCF achieves better precision and F1 scores as compared to other similar approaches merely by using ranked features (these features are less prone to add false positives) at the initial stages and makes use of Research2Vec embedding model to find contextual similarity between two clusters rather than merely relying on proximity measures in traditional approaches. The complexity of MHCF is $O(Bin^3)$, where B = number of blocks, i is the number of features and n is the number of papers. For 100 Blocks, 4 features, and 1550 papers, the worst case took approximately 10 minutes to generate the results. This observation is consistent with the experimental execution time.

MHCF, using five top-ranked features, is scalable across larger datasets. By focusing on features that are neutral with respect to ethnicity, domain agnostic, and which accurately distinguish authorship with minimal false positives, MHCF is able to maintain high precision and F1 scores for author name disambiguation. The complexity of MHCF remains manageable with i capped at 5, allowing it to efficiently handle larger datasets without a significant increase in computational cost.

The experimental results show that MHCF outperforms SAND1 (+31%), SAND2 (+22%), HHC (+32%), MDC (+12%), GFAD (+18%), ATGEP (+32%) and ESMD (+3%) in average pF1 scores. MHCF also performs better than its two variations MHCF-G and MHCF-GL. MHCF-G is a variation of MHCF which uses pre-trained word2vec model trained on the Wikipedia corpus by Google to create a global feature vector. Similarly, MHCF-GL is based on using pre-trained GloVe based word embedding model provided by Stanford on Wikipedia text corpus. MHCF performs better than MHCF-G and MHCF-GL achieving a 5% and 10% gain in pF1 respectively.

In addition to this, it is also shown in the chapter that the overall performance of MHCF can be increased by using useful features instead of using less useful features. A 100% pP and a 98% pF1 score is witnessed merely using author email and author affiliation feature. Whereas, it gives an overall pF1 score of 93.5% with co-authors, author affiliation, paper title, and paper venue feature combinations (it is already discussed in Chapter 3 that paper title and paper venue features are more prone to add false positives in the data).

The next chapter discusses the proposed dataset curation process, which focuses on filling the gaps identified in the literature review process.

Chapter 5

Completing Features for Author Name Disambiguation

The focus of this chapter is to discuss the details of the proposed dataset which addresses the research question (RQ) #2, i.e.

> **How to curate an AND dataset which is: feature enriched, covers multi-disciplinary scholarly data, and encloses authors belonging to multiple ethnic groups?**

The proposed dataset "CustAND" provides a set of 7886 publication records, where each record covers thirteen useful[1] AND features. The dataset has multi-disciplinary publications of authors who belong to multiple Ethnic Groups (EGs), such that the dataset is not skewed with both aspects.

5.1 Introduction

The contribution being discussed in this chapter is an effort to provide a dataset that is feature enriched and includes authors from multiple EGs, working in different domains.

[1] list of useful AND features are identified through literature and are discussed in chapter 2

For this purpose, ambiguous author names are identified from literature but, unlike other datasets, the citation extraction having these names is done using DBLP as well as GS. This results in the inclusion of authors working in diverse domains, belonging to different EGs. The missing feature values in the candidate citations are extracted by going through publication web pages and PDF files. A group of graduate students manually cross-checked the publications and authors' metadata and confirmed authorship linkages using external sources by assessing their affiliations and emails. Though CustAND is limited in scale in contrast to automatically labeled datasets, however, unlike them, it is verified.

Therefore, the rest of the chapter is organized as follows. The section proceeding introduction discusses the complete curation process of CustAND. This is followed by the sections which are: analysis of CustAND, CustAND statistics, and CustAND comparison with existing AND datasets. These sections are later comprehended in the discussion section, which is followed by the novelty of CustAND, after which precedes the chapter summary.

5.2 CustAND Curation Process

The overall process of curating CustAND dataset (as adopted by other similar studies [44]) involves the following major steps:

1. Identify top most ambiguous author names from literature and select candidate ambiguous names such that their are multiple distinct authors with this name.
2. Collect citations with authors names identified in step 1 using DBLP and GS.
3. Find, extract, and annotate missing information in the raw data attained from step 2 using different sources.
4. Pre-process the raw data attained from step 3 to generate csv files for further processing.

5. Manually cross-check and confirm authorship's of the data.

6. Apply Cohen's kappa coefficient (κ) [33, 95] and percent agreement to measure the agreement between the team of annotators confirming authorship's.

Figure 5.1 shows the complete methodology of the data curation process, whereas details of each step are given as follows.

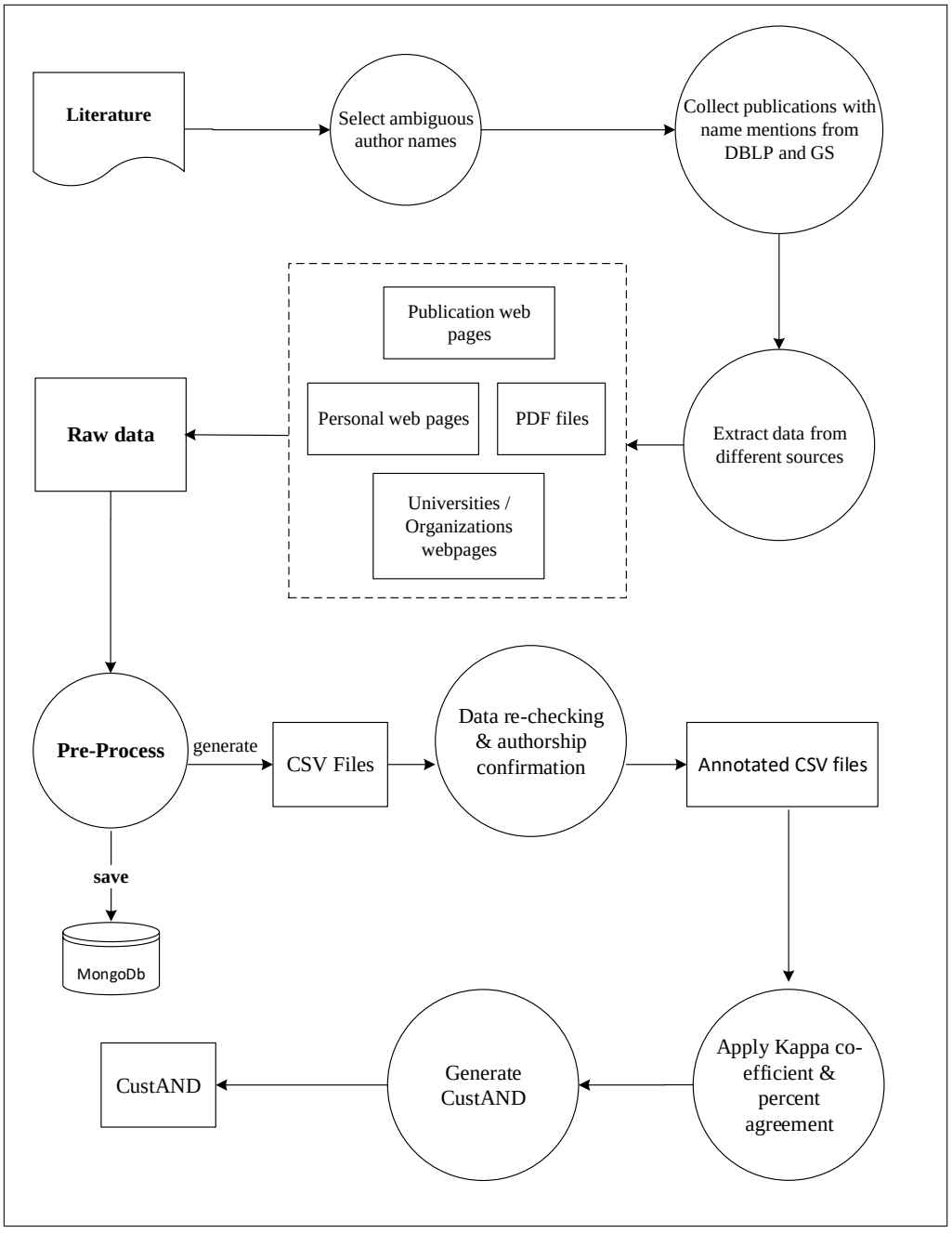

FIGURE 5.1: CustAND Curation Workflow

5.2.1 Identify and Select Ambiguous Author Names

To identify the top most ambiguous author names, a comprehensive literature survey is conducted to identify studies in which researchers have already discovered them. Based on the existing studies [23, 25, 26, 57], a combined candidate list of most ambiguous authors names is made.

The candidate names are manually searched in DBLP and GS to find candidate citations. The results are observed to shortlist such ambiguous authors names citations that seem to belong to different authors, i.e., at least two same name authors or at least five authors sharing any name variant but different emails and affiliations. Based on this criteria, 14 ambiguous author names with 137 distinct authors are finalized.

5.2.2 Citation Collection Sources

After ambiguous author name finalization, candidate author names are searched using DBLP and GS to collect all citations with the particular ambiguous name in it.

DBLP is commonly used by the researchers for AND dataset curation process, and most of the candidate ambiguous names are taken from this platform, therefore, it is used to extract their citations for this study as well. Also, GS is used for this purpose, because, it has attained a lot of popularity due to its scholarly search engine service for multi-domains, which is estimated to roughly contain 389 million documents including articles, citations, and patents, making it the world's largest academic search engine in January 2018 [96]. Moreover, it is also used to include ambiguous authors citations, who work in other domains, than, CS.

Therefore the citations are manually collected from these platforms, which have basic features including co-author names, paper titles, paper publishing venues, and paper publishing year.

5.2.3 Extract and Annotate Missing Information

The raw citations are further manually enriched against missing features, along with more feature values using different sources. They are: publications web pages, publications PDF files, authors different profiles maintained on different platforms, and personal web pages.

5.2.4 Customized Scripts to Process Raw Data

After manual data collection, the data is pre-processed using a customized Python script. After this, the data is saved in MongoDB to generate finalized raw data CSV files. The following processes are performed automatically before proceeding further:

1. Near duplicates (De duplicates) of research papers (using paper title matching) are removed using cosine similarity score with a threshold value of 0.9. (This step is necessary as it is quite possible that multiple citations get included during the manuall collection process.)

2. Standard pre-processing methods are used to clean the raw data which include removing html special characters, non-English characters from author names, removal of brackets, commas, semicolons, and slashes from authors names, using Regular Expressions (RegEx). Similarly, removing non ASCII characters from paper titles, hyperlinks, and conversion of data to lower case is ensured. Finally, 137 CSV files are generated after processing the raw data.

5.2.5 Raw Data Cross Checking and Authorship Confirmations

At this stage, 137 pre-processed CSV data files are disseminated to a team of three graduate students for data re-checking and authorship confirmation purpose.

The process includes, rectification of any error during the manual data extraction process and incorrect authorships. Authors and their respective publications associations are confirmed by looking at the author affiliations and emails metadata. In addition to this, the authorships are also confirmed by searching the authors on ResearchGate (RG) and GS profile pages. Any publication which appears on both profile pages, after confirming author affiliations and emails, confirms their authorship.

The reasons to consider RG and GS for this step is as follows:

1. Profile created by a researcher on RG includes an overview of the researcher, covering his/her skill and expertise, discipline, interests, their full names, list of research papers, figures, and data, etc. To make a profile on RG, the users are required to fill in the necessary information along with their authored publications manually to ensure authenticity of the information.

   Additionally, to remove any other discrepancy and inconsistency, RG automatically contacts co-authors of the researchers' sharing authorships within a publication (if they have a profile on RG) from time to time via emails, asking to confirm whether the publication is authored by the author or not. Therefore the publications appearing under RG profiles are authenticated and updated that too by the authors themselves.

2. GS facilitates researchers to create a profile from existing GS data, which displays their publications and citation information. The researchers fill in the sign-up form entering their full names, affiliations, interests, email addresses, and are requested to verify articles that may have been written by them. Researchers can also search their published articles in GS and add them to their GS profiles. It also calculates the researchers' h-index and i10-index along with information regarding most frequently co-authored researchers etc.

3. Presently, researchers are more focused to use these platforms to interact and collaborate with other researchers as well as showcase their research

contributions instead of going through the hassle of maintaining separate web pages for research display.

Therefore, RG and GS platforms are considered to reconfirm authorships of ambiguous authors and their publications. Any publication appearing on both RG and GS profile pages of an author, and share same affiliation or emails, confirms the citations author authorship. In case of conflict during the data rechecking and authorship confirmation process, percent agreement system is adopted, whereas unresolved cases are dropped from the final dataset.

5.2.6 CustAND Dataset

After manual cross checking and annotation phase, the finalized CSV files are read and κ is applied, which measures the agreement between annotators who each classified N publications into C mutually exclusive groups (distinct authors). The data with κ values between 0.90 and 1.00 are finalized, representing perfect agreement between the annotators. Finally, the finalized data is used to generate "CustAND" using Python script making a tab delimited txt file.

5.3 CustAND Dataset Analysis

5.3.1 Data Records

The complete CustAND dataset is available through github repository with url as follows: (https://github.com/humaira699/CustAND_Full.git).

The dataset consists of one "tab delimited" text file containing all data records, where each record has more than eleven feature values. Each feature value within a record is separated by "/t". Whereas each record within a file is separated by "/n".

Table 5.1. lists all the features available in CustAND.

5.3.2 Technical Validation

The validity of the data is ensured in two steps using the percent agreement evaluation method along with kappa co-efficient κ, by a team of three CS graduate students fully conversant with the domain.

5.3.3 First Step

In the first step of data validation, the annotators are required to check the validity of the raw data which is manually collected from multiple data sources. For this, each student is given the raw data which is in the form of CSV file i.e., 137 files are given to each annotator, where each file represents a distinct author. The team cross-check each CSV file by searching the publication's web pages and pdf files to rectify any discrepancies. The annotators compare each file and remove errors if present. They then exchange their files with each other and rate them with either 0 or 1, to show agreement or disagreement i.e., whether the file holds the correct data with respect to the publications web pages/pdf files or not.

Following this procedure, all three annotators rated the data showing 100% agreement[2].

5.3.4 Second Step

In the second step of technical validation, the team is required to manually confirm the author and their authored publications which are present in each CSV file. Each annotator confirms this by looking at the author affiliations, email, and data presence on RG and GS profiles of the authors. This entire activity ensures publication authorships. After confirmation, the annotator's data are confirmed with each other using Cohen's Kappa co-efficient κ. κ is a statistical method that is commonly used to do interrater testing. Its score can range from -1 to +1, where

[2] A.27, of Appendix A

0 represents the amount of agreement that can be expected from random chance and 1 represents perfect agreement between the raters.

TABLE 5.1: CustAND Feature's Description.

#	**Features**	**Records with values**	**Multi values**	**Multi values separator symbol**	**Short description**
1	Block id	7886	×	None	Unique id to identify ambiguous author name groups
2	Author id	7886	×	None	Unique id to identify distinct authors within an ambiguous group
3	Paper id	7886	×	none	Unique id to identify papers within the dataset
4	Author full name1	7886	×	none	Author in question full name mentioned on his/her GS profile page
5	Author full name2	7886	×	none	Author in question full name mentioned on his/her RG profile page. (Two names can be slightly different on RG and GS profile pages e.g., "m a a shoukat choudhury" (mentioned on GS profile) and shoukat choudhury (mentioned on RG profile) of the same author).

6	Author affiliation	7886	×	none	Author affiliation mentioned on his/her profile page
7	Author email	7591	×	none	Author email domain mentioned on his/her profile page e.g., cust.edu.pk
8	Author research interests	7340	✓	,	Author research interests mentioned on his/her profile page e.g., "data mining, machine learning, information retrieval" etc.
9	Co-authors name	7886	✓	;	Author and co-authors name who authored the paper e.g., Humaira Liaquat; M Abdul Qadir
10	Paper title	7886	×	none	Title of the paper
11	Paper publishing venue	7246	×	none	Venue of the published papers
12	Author affiliation	6160	✓	;	Author in question affiliation mentioned in the paper. Missing value is replaced with "none" string
13	Author email	1029	✓	;	Author in question email mentioned in the paper
14	Paper abstract	5296	×	none	Abstract of the paper
15	Paper keywords	3257	✓	,	Keywords of the paper

16	Paper publishing year	7867	×	none	Paper publishing year

The calculation of Cohen's kappa is performed using the formulas given by [33]. Whereas, Kappa result interpretations[3] used in this study are listed by [95], and covered in detail in Appendix A. The scores show that the curated data is in almost perfect agreement with all the annotators [4].

In summary, we can say that the finalized data is not erroneous and contains the correct authored data with respect to their authors.

5.4 CustAND Statistics

This section covers the CustAND statistics which include the following aspects:

1. An overview of **CustAND specification**, along with their description is given. It covers the general details of the dataset. Refer to Table 5.2 for details.

2. Next, the complete **statistics of the ambiguous authors count** is given in Table 5.3, which covers:

   (a) Number of distinct authors sharing the same full names.

   (b) Number of distinct authors sharing name variants.

   (c) Number of phonetically same named authors.

   (d) Ethnicity distribution of the authors. Similarly, it is seen that the data in CustAND is not skewed [5].

[3] A.28

[4] Kappa statistic calculation with respect to the data can be found in Appendix A, Table A.15, A.16, and A.17 representing data recorded against annotator 1 with annotator 2, annotator 1 with 3, and annotator 2 with 3 respectively, and equation A.1, A.2, and A.3 (Appendix A)

[5] B.3 of Appendix B

3. The next subsection gives an **overview of the publications associations with respect to ambiguous authors, along with the co-authors distribution with respect to publications in CustAND**. Refer to Figure 5.2.

4. **Domain**[6] **Distribution** of data in CustAND.

5. Next, the number of citations (instances) per ambiguous authors without missing values against some commonly used feature combinations in the literature. Refer to Table 5.4 for this insight.

6. Miscellaneous subsection covers details of the year-wise publications and the number of publications with respect to distinct authors (with the same full names).

5.4.1 CustAND Specification

The following Table 5.2 covers the CustAND specifications.

TABLE 5.2: Specification and Description of CustAND.

Specification	**Description**
Subject	Computer Science
Purpose of the dataset	Develop and test author name ambiguity problem resolving techniques using this dataset.
Type	Research papers and their author's metadata (text file).
Total authors	137
Total ambiguous blocks	14
Total records	7886

[6]Refer to Appendix B, Figure B.4 for its distribution graph.

Annotators	3
Data extraction year	2019
Experimental factors	Standard text pre-processing methods are applied.
Experimental features	Raw data extracted manually, which is rechecked and further annotated by a team of graduate students.
Data source location	Publication websites and pdf files, authors personal web pages, and different profile pages

5.4.2 CustAND Authors Statistics

Table 5.3 lists the complete statistics of the authors in the dataset.

TABLE 5.3: Statistics of CustAND.

Names	**Distinct authors**	**Total publications**	**Same name**	**Same phonetic name**	**Same name variant**
A Choudhary	12	548	0	0	12
M A Qadir	15	354	0	4	11
A Gupta	8	878	6	0	2
A Kumar	9	191	2	0	7
Bin Li	8	436	5	0	3
D Eppstein	3	36	2	0	1
J Lee	8	444	0	0	8
J Martin	9	708	2	0	7
J Mitchell	10	766	5	0	5
J Robinson	12	409	5	0	7
J Smith	12	511	5	0	7

S Kim	10	420	4	0	6
Z Zhang	10	978	0	0	10
K Tanaka	11	1207	2	0	9
		Ethnical groups and number of authors.			
Arab	19	Chinese	18	English	37
French	6	Hispanic	6	Indian	22
Japanese	12	Korean	17		
Mean	17	**Standard Deviation**	10		

The EG distribution[7] shows that the dataset is balanced[8] distributed up to three standard deviations.

5.4.3 Author Distribution per Publication and Publication Distribution per Ambiguous Author

Figure 5.2 (a) shows that CustAND includes 9% single and 15% double authored publication records. Whereas, it includes 58% publications with more than two and less than six authors in it. Also, 14% of the publications have co-authors count between six and ten. Whereas, 4% of the publications have more than 10 and less than 101 co-authors. Such scenarios are helpful to design solutions that can work in case of no co-authors, less number of co-authors, and many co-authors (as in those cases the probability of having ambiguous co-authors themselves is more).

Figure 5.2 (b) shows the publications distribution per author in CustAND. It is observed that 21% of authors have publications ranging between one and ten. Also, 21% of the authors have publications ranging between eleven and twenty, 42% of

[7]Figure B.3, Appendix B

[8]In a study by D. G. Altman [97], it is highlighted that a normal distribution extends beyond two standard deviations on either side of the mean.

the authors have more than twenty, and, less than 101 publications associated with them.

Inclusion of such data helps to design techniques that work well in real-world scenarios. For example, there are different types of researchers with respect to their research activities. Some researchers are active and have more publications to their credit. Similarly, some researchers are less active and have very few publications to their credit. Also, there can be young researchers, who recently started their research career, therefore they too have less publications to their credit.

The distribution given in the figure points out that CustAND includes such data in it.

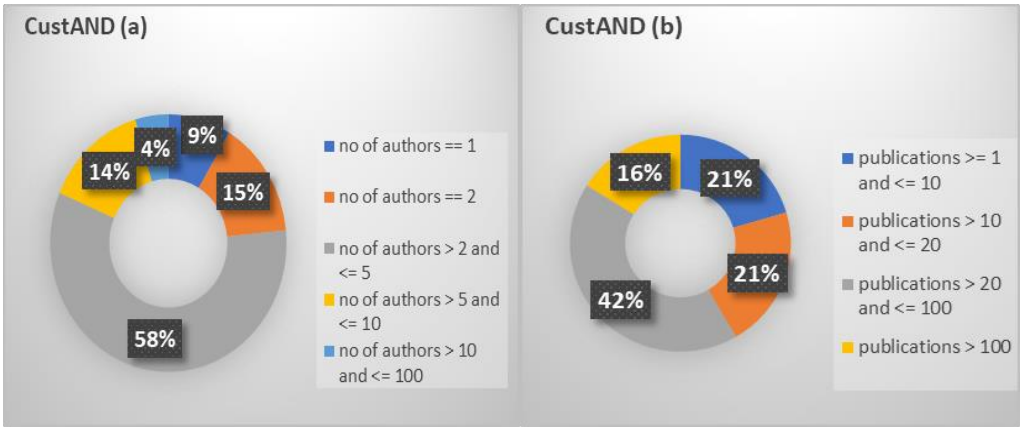

FIGURE 5.2: Distribution of Co-authors per Publication and Publication Records per Ambiguous Author in CustAND Collection.

5.4.4 CustAND Domain Distribution

As far as the domain distribution[9] of CustAND is concerned, each author's research interests and paper titles are manually analyzed and categorized as per Pakistan Higher Education Commission (HEC) approved list of disciplines and subjects (HEC, n.d.). This study broadly categorized each author in one of the three domains.

[9]Figure B.4 of Appendix B shows that CustAND is normally distributed.

1. **Bio Sciences and Medicine Related Research Areas**, which cover domains like biology, chemistry, life sciences, physiological sciences, medicine, etc.

2. **The Social Sciences** domain which includes areas such as anthropology, business & management, human geography, law, media studies, political science and international relations, social policy, sociology, library, and language related areas.

3. **The Engineering Sciences** domain covers mechanical engineering, nanotechnology, physics, quantum theory, electronics, electrical engineering, and computer science domains.

5.4.5 CustAND Instance Count Against Common Features

Table 5.4 highlights the count of records without missing values, per ambiguous author name against different commonly used feature combinations.

TABLE 5.4: Instance Count Without Missing Values per Ambiguous Author per Specified Feature Combinations.

Blocks	**(Co-authors, Title, Venue)**	**(Title, Venue, Affiliation, Year)**	**(Title, Venue, Email)**
A Choudhary	543	540	540
M A Qadir	337	314	314
A Gupta	846	827	829
A Kumar	188	188	188
Bin Li	436	436	436
D Eppstein	35	35	35
J Lee	433	433	433

J Martin	700	700	700
J Mitchell	738	556	556
J Robinson	403	403	403
J Smith	500	499	499
S Kim	418	415	416
Z Zhang	699	315	314
K Tanaka	978	416	416
Total	**7254**	**6077**	**6079**

5.4.6 Miscellaneous

CustAND[10] includes a larger number of publication records which are published between the years 2013 and 2019, whereas a minimum number of instances are included in the dataset which are published before the year 1986.

CustAND either includes almost the same number of publications with respect to distinct authors or has variations in the publication counts[11]. For example: a distinct author named 'Anik Kumar Gupta' has publications instance distribution as; 60, 84, 108, etc. Some authors in same-named-author groups have varied publication record counts. This points towards the fact that the dataset includes diverse scenarios which will be helpful in designing generalized solutions for AND.

5.5 CustAND Comparison

This section covers a comparison of CustAND with nine publicly available datasets, which are reviewed in Chapter 2. The comparison is done based on the gaps which are identified in "dataset analysis section (2.2.2.2)". They are:

[10]Appendix B, Figure B.5 shows the number of papers published per year that are included within this dataset irrespective of the author to whom they belong.

[11]Figure B.6 of Appendix B show the number of publications with respect to distinct authors

1. Useful features coverage
2. Ethnicity of the authors
3. Domain of the publications included in the dataset
4. Labeling strategy of the dataset

From Table 5.5, it can be analyzed that most of the reviewed datasets (Chapter 2) are hand labeled and their data is manually validated. Similarly, these datasets are mostly domain-specific, i.e., except PubMed, Medline, scadZBMATH, and Inspire, remaining datasets are CS domain specific. Whereas PubMed and Medline are based on medicine related field, Inspire is curated from physics domain and scadZBMATH is based on Maths domain. As far as the EGs are concerned, the analyzed datasets are skewed with this respect as well. The analysis related to this angle is covered in Chapter 2.

In comparison to the reviewed datasets, CustAND follows a hybrid approach to label the data. Where the problem of inclusion of false positives in the hand labeling strategy of the data is addressed by analyzing and considering multiple external profiles simultaneously along with the publications metadata. The ambiguous author's publications are extracted using GS as well as DBLP, due to which authors of different EGs, working in multiple domains are included in CustAND. Similarly, as discussed previously, the dataset is not skewed with respect to the author's EGs and the domain of the publications which are included in the dataset. Similarly, the useful feature coverage in CustAND is better as compared to all the reviewed datasets. The statistics related to feature coverage in CustAND are already given in Table 5.1, under section 5.3.1.

TABLE 5.5: Comparison of CustAND Reviewed Datasets.

No	Dataset	Domain	Features	LStrategy	Ref	EG distribution
1	**CustAND**	**Multi**	**13**	**Hybrid**	[15]	**Normal**

2	DBLP	CS	4	Manual	[39, 57]	Skewed
3	BDBComp	CS	5	Manual	[20]	Skewed
4	Arnetminer	CS	6	Manual	[25]	Skewed
5	KISTI-AD-E-01	CS	5	Manual	[26]	Skewed
6	PubMed	Medical	1	Manual	[44]	Skewed
7	Aminer	CS	7	Manual	[9]	Skewed
8	Medline	Medical	1	Manual	[50]	Skewed
9	Pubmed	Medical	11	Semi-Automatic	[74]	Skewed
10	Medline	Medical	11	Semi-Automatic	[74]	Skewed

5.6 Discussion

Findings of this study can be summarized as follows: every dataset labeling strategy has its own set of pros and cons. Unlike others, CustAND follows a hybrid approach to label the data, which counters the inclusion of false positives in the dataset, due to the unavailability of information. This is done by considering multiple profiles of authors simultaneously to ensure data authenticity. In case of doubt, the candidate author or publication record is dropped altogether.

General analysis of the evaluated datasets shows that the reviewed datasets are domain-specific, whereas the authors with respect to the EGs are skewed (refer to Chapter 2 for this gap analysis).

In CustAND, the author who is being disambiguated and their co-author names are maintained in full form, along with their names mentioned on different profile pages (which sometimes vary). It contains 9% single authored publications along

with 15% of the publications which have a single co-author, making it up to 24%. Though this number is better than most of the evaluated datasets but can be improved further in the future.

CustAND distributes its ambiguous authors names in three major domains, it holds eight EGs, it includes more than eleven useful features such that, 13% of the author emails are included in the dataset, 41% paper keywords, 67% paper abstracts, 78% of author affiliations, 92% paper publishing venue, 93% author research interests as mentioned on his/her profile page, 96% of author email domains (extracted through author's profile pages/personal web pages), 100% author in question full name mentioned on his/her profile web page 1, 100% author in question full name mentioned on his/her profile web page 2, 100% author full name, 100% co-author full names, 100% author sequence is maintained, 100% paper titles, 100% paper publishing year and 100% author affiliations are maintained on his/her profile pages.

The data is gathered from multiple web sources i.e., author affiliation is gathered from his/her personal web page vs the affiliation present in his/her authored publication. Similarly, his/her email is mentioned on the personal web page vs the email given within the publication. The feature values can be different i.e., emails and affiliations appearing in publications and which are mentioned on his/her profile pages. This can be an indicator showing the change in emails and affiliations of an author due to the change of jobs. The data is present in the dataset from two perspectives i.e., author information coming from publications metadata and author information coming from the author's personal web page. This data will be helpful to design author name ambiguity resolving techniques from both author assignment and author grouping perspectives. Author research interests gathered from personal web pages of distinct authors give a broader and generic view of topics that can be used to develop and test techniques following topic modeling.

The unavailability of some feature values against some instances in CustAND is often due to their absence in the publication file or because of limited access to publication PDF files. Author email feature value presence is the least among

all others because of its absence from the publication metadata (as the author in question is not the corresponding author, or, the email data is absent altogether, or, due to the limited access of the publication PDF file). Almost similar problem is faced while extracting author affiliation and paper keywords feature values. As far as remaining useful features that are not covered by CustAND are concerned, it is done intentionally. This is because, to the best of knowledge, only one literature evidence can be found that graded these features as useful [42], whereas, other researchers have either not assessed their usefulness or have not declared them among the top useful features. Additionally, the "Address" feature, listed as one of a useful feature is not maintained separately in CustAND, rather the affiliation feature is a combination of the author designation, postal address, city, and country values.

To see the effect of the use of useful feature enriched labeled dataset by an AND technique in comparison to the datasets which are formulated specifically with the perspective of limited information availability, an experiment is conducted, in which the proposed AND technique MHCF [34](chapter 4), proved to perform much better using only two but useful features in comparison to a feature combination which itself is prone to cause ambiguity.

Therefore, it is a view that the datasets should be enriched with a reasonable amount of information which have more discriminating powers to discriminate distinct ambiguous authors from one another [15]. The dataset should include multiple domains data, as pblications of one domain may not have complex patterns in it which cause challenging scenarios. For example, usually medicine domain publications have many co-authors in them, Due to this the probability of encountering ambiguous co-author names become higher, which makes the co-author feature more error prone in cases where the number of co-authors are less.

Therefore, it is concluded that it is perfectly reasonable to disregard and use limited information while developing an author name ambiguity resolving technique but should be an option given to the researchers rather than limiting the datasets with eliminating necessary dimensions in it.

5.7 Novelty of CustAND

The novelty of the CustAND dataset is evidenced through the following key contributions:

1. Unlike existing datasets in AND domain, CustAND is characterized as having feature-rich composition, encompassing more than eleven better result producing features. These features have been carefully selected based on their positive impact on the authorship results (Chapter 3 covers this aspect). The inclusion of these features in CustAND elevates the utility of CustAND, surpassing the capabilities of other datasets in terms of feature diversity.

2. CustAND addresses the limitations of existing AND datasets, which are often domain-specific, focusing solely on disciplines like Computer Science, Medicine, Physics, or Mathematics. By including data from three major academic domains—Bio-Sciences, Social Sciences, and Engineering Sciences, CustAND broadens its usage to develop AND techniques. Further, CustAND includes authors from eight distinct ethnic backgrounds. This diversity enhances the dataset's utility for AND research.

3. CustAND includes 14 ambiguous author names, representing 137 distinct authors. This carefully curated collection of ambiguous names will facilitate in testing and evaluation of AND techniques, offering researchers to assess AND techniques under varying scenarios and complexities.

5.8 Chapter Summary

This chapter encloses details regarding the proposed dataset CustAND, covering its data curation process, the validity of the data, its specifications, and comparison with the reviewed datasets. CustAND, unlike the reviewed datasets, is feature enriched, which covers more than eleven useful features (such that the percentage

of their values is better than the reviewed techniques), that are proven to impact the authorship results in a positive manner (refer to Chapter 3). Moreover, CustAND, Unlike other datasets, is comprised of multi-domain data (covers three major domains i.e. BioSciences, Social Sciences, and Engineering Sciences) where the authors included in the datasets belong to multiple ethnic groups (eight ethnic groups). Therefore, CustAND provides a set of 14 ambiguous author names in total with 137 distinct authors.

To curate CustAND collection, the raw data is collected from DBLP and GS, which is later annotated, checked, and confirmed using different data sources by a team of graduate students with 100% agreement and Cohen's kappa coefficient score ranging between 0.95 and 1. Similarly, the validity of the data is ensured by tallying the author's profiles maintained on different platforms, more specifically GS and RG.

CustAND, besides holding feature-enriched, multi-domain, and multi-EGs data, also has a better percentage of single-authored publications included in it. It also records multiple affiliation and email values against these features. Similarly, the percentage of availability of feature values is better than the reviewed techniques. However, the absence of feature values against some feature instances is either due to the limited access to the publication data or its absence from the source.

The implications of these aspects in the data will allow the development of such AND techniques that can cater to more demanding scenarios. Like: 1) single-authored publications employ challenges to AND techniques. The reviewed datasets hold a very low percentage of such scenarios in them, restricting the testing of AND techniques with this regard [14]. 2) preserving multi-affiliation and email values in CustAND will help develop techniques that fail in cases when these values change due to the switching of jobs. Above all, CustAND is curated to achieve the following: 1) study the impact of useful features on the author name disambiguation process, as the reviewed datasets are curated for feature-scarce scenarios. 2) After developing an enhanced AND technique, test it on the proposed dataset to see the effect of the use of useful features. 3) Provide the research community

with a dataset that is feature-enriched and diverse with respect to data domains and EGs. Because, information should be made available to researchers, allowing them to decide how to use it, rather than constraining datasets from holding and providing this information.

Chapter 6

Conclusion and Future Directions

6.1 Conclusion

Digital libraries and different scholarly data search engines index and make available lists of thousands of scholarly articles against different authors. It is commonly known that this process is susceptible to errors, as, the authors often have similarities in their names. The name similarity can be in the form of sharing common names (homonyms) or may share a variation of their name, i.e. when the authors publish under different name variations (synonyms) [1] (often termed as author name ambiguity problem). Since the research fields are facing a rapid increase in scholars and their publications, author name ambiguity has become an inevitable problem, and digital libraries and scholarly search engines often fall prey to false academic authorships. This leads to incorrect assessments of researchers' research worth, often required by different organizations and universities to scrutinize researchers for award assignments, hiring purposes, and research funding assignments.

Due to the increasing number of researchers, the author name ambiguity problem is alleviating at a rapid pace, and requires, automatic author name disambiguation techniques to improve academic authorship results.

Many techniques have been proposed for author name disambiguation, which uses supervised, unsupervised, and graph-based learning models, but, most of them suffer from low precision, recall, and F1 scores. When precision is improved, it usually comes at the expense of lower recall, and vice versa. This trade-off affects the overall F1 score of the technique. Majority of the existing AND techniques overall results (F1 scores) are reported to be distributed between the range of 66-77%, and a few between the range of 88-99%. However, techniques with higher scores mainly use ethnic and domain-centric datasets, which are skewed in these aspects.

To improve the overall result (F1 score) of an AND technique, the identification and usage of appropriate features is an important aspect. This proposed study insights that very few studies have evaluated the effects of features and those who did have only considered a subset. Also, incomplete knowledge is present in the literature regarding better F1 score-achieving feature combinations. It is also observed in the literature review process that some studies rate certain features as low, whereas, others rate the same feature as high. In order to address the gaps, the proposed study develops a feature ranking scheme, with an outcome of the proposed feature ranking. Also, in this study, useful feature combinations are identified which gives better F1 scores without compromising the precision.

To achieve the proposed feature ranking and identify better F1 score-producing feature combinations, first, a detailed review of the features availability in existing publicly available AND datasets is done. Also, the domain of the publications and the ethnicity of the authors in the reviewed datasets are identified. It is concluded that majority of the publicly available AND datasets provide limited useful feature coverage. They are mostly curated keeping in mind particular scenarios and feature availability's. It is also observed that these datasets are domain as well as ethnic-centric. This means that they are skewed with respect to the domain and ethnic groups of the authors. The in-availability of these aspects makes the datasets specific, which ultimately limits the author name disambiguation techniques capability to address the problem in real world. Therefore, the proposed study includes details regarding the proposed AND dataset (CustAND)

which is publicly available to the research community. The dataset includes thirteen useful features, which is more than the reviewed datasets, that too with higher percentages at the instance level. Unlike existing datasets, CustAND is not domain-centric, rather, it includes publications data belonging to biomedicine, engineering, and social sciences domains. Similarly, the ethnicity of authors is not region or ethnic-specific. Rather, the authors included in the dataset are skewed with respect to authors ethnicity, as well as domains. This study also includes the experimentation and results of using the proposed AND dataset using the proposed author name disambiguation technique. The results conclude that the authorship results improve by using useful features rather than using any available feature.

To improve the authorships of authors without compromising the precision scores, this study proposes a multi-layer heuristics based author name disambiguation technique, which uses set of proposed ranked features and combinations. The proposed technique is evaluated using different datasets against a number of techniques including word embedding-based approaches, heuristics-based approaches, graph-based approaches, and hybrid approaches. The proposed technique results clearly show that it is better than the existing techniques in terms of their authorship results (F1 scores) without compromising the precision.

6.2 Novelty and Contribution of the Research

The novelty of this research is embodied in the development of the Multilayer Heuristic Based Clustering Framework (MHCF), which integrates several innovative components to enhance the effectiveness of Author Name Disambiguation (AND):

1. **Novel Feature Ranking Scheme**

   (a) **Initial Feature Ranking from Literature Review:** The research begins with a literature-driven approach to assign preliminary ranks

to candidate features. These initial rankings are then refined through experimental validation, ensuring a thorough and informed feature selection process.

(b) **Optimized Feature Combinations:** The methodology identifies feature combinations that enhance precision, recall, and F1 scores, contributing to the overall improvement in the accuracy of AND techniques.

(c) **Experimental Validation:** The proposed feature rankings undergo rigorous experimental validation across multiple datasets to ensure their effectiveness. This validation process mitigates the risk of inconsistencies, providing confidence in the reliability of the rankings.

(d) **Comparative Analysis with Existing Rankings:** A critical comparison of the proposed feature rankings with existing methodologies offers a clear and concise perspective, demonstrating feature contribution to enhance AND.

2. **Novel AND Technique (MHCF)**

(a) The MHCF technique introduces a novel approach to AND by incrementally utilizing ranked features, complemented by intelligently designed rules for clustering academic authorships. This approach significantly reduces false positives, enhancing the accuracy of the disambiguation process.

(b) MHCF integrates the Research2Vec model, which is trained on the arXiv dataset, to generate semantically coherent vectors that reduce false positives and improve the recall and precision as compared to other pre-trained embedding models.

(c) Comprehensive experiments demonstrate that MHCF achieves substantial improvements in precision, recall, and F1 scores compared to other existing AND techniques.

3. **CustAND Dataset**

(a) Unlike existing datasets, CustAND is feature enriched, encompassing over eleven impactful features, and holds data from three major academic domains, including Bio-Sciences, Social Sciences, and Engineering Sciences. This broadens its applicability in the AND domain.

(b) CustAND includes authors from eight distinct ethnic backgrounds, enriching the dataset's representational breadth for developing and testing AND techniques.

(c) CustAND includes a set of 14 ambiguous author names, representing 137 distinct authors, for testing and evaluating AND techniques under varying complexities and scenarios.

The comprehensive framework, combining the MHCF approach with the Research2Vec model and the CustAND dataset, offers a significant advancement in author name disambiguation, addressing critical gaps in existing methods and enhancing the accuracy and applicability of AND techniques.

6.3 Implications of the Proposed Research

The proposed study carries significant implications for both the academic and practical domains.

1. **Improved Accuracy of Authorships:** The proposed technique and its methodology significantly enhance the accuracy of academic authorships. This will streamline the process of identifying and attributing research contributions to the correct authors, thereby improving the reliability and usability of academic search engines and digital libraries. This will greatly benefit researchers, students, and professionals who rely on these resources for their work.

2. **Contribution Towards Reliable Bibliometric Indicators:** Enhancement in the author name disambiguation process is quite crucial for conducting comprehensive bibliometric analyses. The proposed technique will

facilitate the calculation of reliable bibliometric indicators and metrics, by giving better authorship results, thus revealing the true scholarly impact and productivity of a researcher.

3. **Promotion of Collaboration:** By enhancing authorship results (F1 scores) with better precision, this study will lead to better research collaborations among researchers. Institutions, funding agencies, and journals can better identify potential collaborators and allocate resources effectively to foster interdisciplinary research.

6.4 Future Directions

In the future work, a list of tasks can be considered. The list is not complete and there may be other tasks that can be considered as well.

1. In this study, a heuristics-based author name disambiguation approach MHCF is developed, which shows improvement in authorships as compared to the existing AND techniques. However, there is still room for improvement which can be done in our future work.

2. Another direction that can be explored in the future is to make MHCF work in online mode, such that its precision, recall, and F1 scores are better than the existing techniques working on this line. Because, the existing techniques working in online mode are suffering from low precision, recall, and F1 scores as well.

3. Continuing with the direction to explore existing AND techniques that work in online mode, a comprehensive literature review also needs to be done, which highlights the problems in them besides their low authorship results.

These tasks are related to our research and cover a range of areas we want to explore in the future.

Bibliography

[1] J. Kim and J. Kim, "Effect of forename string on author name disambiguation," *Journal of the Association for Information Science and Technology*, vol. 71, no. 7, pp. 839–855, 2020.

[2] B. S. Frey and K. Rost, "Do rankings reflect research quality?" *Journal of Applied Economics*, vol. 13, no. 1, pp. 1–38, 2010.

[3] P. Weingart, "Impact of bibliometrics upon the science system: Inadvertent consequences?" *Scientometrics*, vol. 62, no. 1, pp. 117–131, 2005.

[4] M. Färber and L. Ao, "Enhancing the Microsoft Academic Knowledge Graph via Author Name Disambiguation, Publication Classification, and Embeddings," *Semantic Web*, 2020.

[5] M. S. Kumar and P. Neelima, "Design and implementation of scalable, fully distributed web crawler for a web search engine," *International Journal of Computer Applications*, vol. 15, no. 7, pp. 8–13, 2011.

[6] S. Rohatgi, *Design and Data Mining Techniques for Large-Scale Scholarly Digital Libraries and Search Engines*. The Pennsylvania State University, 2023.

[7] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National academy of Sciences*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

[8] K. M. Pooja, S. Mondal, and J. Chandra, "Exploiting similarities across multiple dimensions for author name disambiguation," *Scientometrics*, vol. 126, no. 9, pp. 7525–7560, 2021.

[9] Y. Zhang, F. Zhang, P. Yao, and J. Tang, "Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop." in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018, pp. 1002–1011.

[10] Y. Liu, W. Li, Z. Huang, and Q. Fang, "A Fast Method Based on Multiple Clustering for Name Disambiguation in Bibliographic Citations," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 3, pp. 634–644, mar 2015. [Online]. Available: https://doi.org/10.1002/asi.23183

[11] B. Chen, J. Zhang, J. Tang, L. Cai, Z. Wang, S. Zhao, H. Chen, and C. Li, "Conna: Addressing name disambiguation on the fly," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[12] J.-W. Seol, S.-H. Lee, and K.-Y. Kim, "Author disambiguation using co-author network and supervised learning approach in scholarly data," *International Journal of Software Engineering and Its Applications*, vol. 10, no. 4, pp. 73–82, 2016.

[13] L. Peng, S. Shen, J. Xu, Y. Fu, D. Li, and A. L. Jia, "Diting: An Author Disambiguation Method Based on Network Representation Learning," *IEEE Access*, vol. 7, pp. 135 539–135 555, 2019.

[14] M.-C. Müller, F. Reitz, and N. Roy, "Data sets for author name disambiguation: an empirical analysis and a new resource," *Scientometrics*, vol. 111, no. 3, pp. 1467–1500, jun 2017. [Online]. Available: https://doi.org/10.1007/s11192-017-2363-5

[15] H. Waqas and A. Qadir, "Completing features for author name disambiguation (AND): an empirical analysis," *Scientometrics*, pp. 1–25, 2022.

[16] J. Kim, J. Kim, and J. Owen-Smith, "Ethnicity-based name partitioning for author name disambiguation using supervised machine learning," *Journal of the Association for Information Science and Technology*, 2021.

[17] D. Shin, T. Kim, J. Choi, and J. Kim, "Author name disambiguation using a graph model with node splitting and merging based on bibliographic information," *Scientometrics*, vol. 100, no. 1, pp. 15–50, jul 2014. [Online]. Available: https://doi.org/10.1007/s11192-014-1289-4

[18] J. Zhu, X. Wu, X. Lin, C. Huang, G. P. Fung, and Y. Tang, "A Novel Multiple Layers Name Disambiguation Framework for Digital Libraries Using Dynamic Clustering," *Scientometrics*, vol. 114, no. 3, pp. 781–794, mar 2018. [Online]. Available: https://doi.org/10.1007/s11192-017-2611-8

[19] P. Km, S. Mondal, and J. Chandra, "A Graph Combination With Edge Pruning-Based Approach for Author Name Disambiguation," *Journal of the Association for Information Science and Technology*, vol. 71, no. 1, pp. 69–83, 2020.

[20] R. G. Cota, A. A. Ferreira, C. Nascimento, M. A. Gonçalves, and A. H. F. Laender, "An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations," *Journal of the Association for Information Science and Technology*, vol. 61, no. 9, pp. 1853–1870, 2010.

[21] A. A. Ferreira, A. Veloso, M. A. Gonçalves, and A. H. F. Laender, "Effective Self-training Author Name Disambiguation in Scholarly Digital Libraries," in *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, ser. JCDL 10. New York, NY, USA: ACM, 2010, pp. 39–48. [Online]. Available: http://doi.acm.org/10.1145/1816123.1816130

[22] A. A. Ferreira, A. Veloso, M. A. Gonçalves, and A. H. Laender, "Self-training author name disambiguation for information scarce scenarios," *Journal of the Association for Information Science and Technology*, vol. 65, no. 6, pp. 1257–1278, 2014.

[23] H. Han, W. Xu, H. Zha, and C. L. Giles, "A Hierarchical Naive Bayes Mixture Model for Name Disambiguation in Author Citations," in *Proceedings of the 2005 ACM Symposium on Applied Computing*, ser. SAC 05. New York, NY, USA: ACM, 2005, pp. 1065–1069. [Online]. Available: http://doi.acm.org/10.1145/1066677.1066920

[24] A. Culotta, P. Kanani, R. Hall, M. Wick, and A. McCallum, "Author disambiguation using error-driven machine learning with a ranking loss function," in *Sixth International Workshop on Information Integration on the Web (IIWeb-07), Vancouver, Canada*, 2007.

[25] X. Wang, J. Tang, H. Cheng, and P. S. Yu, "ADANA: Active Name Disambiguation," *2011 IEEE 11th International Conference on Data Mining*, pp. 794–803, 2011.

[26] I.-S. Kang, P. Kim, S. Lee, H. Jung, and B.-J. You, "Construction of a Large-scale Test Set for Author Disambiguation," *Inf. Process. Manage.*, vol. 47, no. 3, pp. 452–465, may 2011. [Online]. Available: http://dx.doi.org/10.1016/j.ipm.2010.10.001

[27] Y. Qian, Q. Zheng, T. Sakai, J. Ye, and J. Liu, "Dynamic Author Name Disambiguation for Growing Digital Libraries," *Inf. Retr.*, vol. 18, no. 5, pp. 379–412, oct 2015. [Online]. Available: http://dx.doi.org/10.1007/s10791-015-9261-3

[28] SourceMedia, "DMReview. Glossary," 2007. [Online]. Available: www.dmreview.com/glossary/a.html

[29] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of management information systems*, vol. 24, no. 3, pp. 45–77, 2007.

[30] A. A. Ferreira, M. A. Gonçalves, and A. H. F. Laender, "Automatic Disambiguation of Author Names in Bibliographic Repositories," *Synthesis Lectures on Information Concepts, Retrieval, and Services*, vol. 12, no. 1, pp. 1–146, 2020.

[31] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[32] S. Alelyani, J. Tang, and H. Liu, "Feature selection for clustering: a review." *Data clustering: algorithms and applications*, vol. 29, no. 1, 2013.

[33] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[34] H. Waqas and M. A. Qadir, "Multilayer heuristics based clustering framework (MHCF) for author name disambiguation," *Scientometrics*, pp. 1–42, 2021.

[35] B. A. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele University and Durham University Joint Report, Tech. Rep. EBSE 2007-001, 07 2007. [Online]. Available: https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf

[36] S. Milano, M. Taddeo, and L. Floridi, "Recommender systems and their ethical challenges," *Ai & Society*, vol. 35, pp. 957–967, 2020.

[37] S. Raj, A. K. Sahoo, and C. Pradhan, "Privacy preserving in collaborative filtering based recommender system: a systematic literature review," *Progress in Computing, Analytics and Networking: Proceedings of ICCAN 2019*, pp. 513–522, 2020.

[38] I. Elnabarawy, W. Jiang, and D. C. Wunsch II, "Survey of privacy-preserving collaborative filtering," *arXiv preprint arXiv:2003.08343*, 2020.

[39] H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis, "Two Supervised Learning Approaches for Name Disambiguation in Author Citations," in *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, ser. JCDL 04. New York, NY, USA: ACM, 2004, pp. 296–305. [Online]. Available: http://doi.acm.org/10.1145/996350.996419

[40] P. Treeratpituk and C. L. Giles, "Disambiguating Authors in Academic Publications Using Random Forests," in *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, ser. JCDL 09. New York, NY, USA: ACM, 2009, pp. 39–48. [Online]. Available: http://doi.acm.org/10.1145/1555400.1555408

[41] J. Wang, K. Berzins, D. Hicks, J. Melkers, F. Xiao, and D. Pinheiro, "A boosted-trees method for name disambiguation," *Scientometrics*, vol. 93, no. 2, pp. 391–411, nov 2012. [Online]. Available: https://doi.org/10.1007/s11192-012-0681-1

[42] M. Levin, S. Krawczyk, S. Bethard, and D. Jurafsky, "Citation-based Bootstrapping for Large-scale Author Disambiguation," *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 5, pp. 1030–1047, may 2012. [Online]. Available: http://dx.doi.org/10.1002/asi.22621

[43] H. N. Tran, T. Huynh, and T. Do, "Author Name Disambiguation by Using Deep Neural Network," *CoRR*, vol. abs/1502.0, 2015. [Online]. Available: http://arxiv.org/abs/1502.08030

[44] M. Song, E. H.-J. Kim, and H. J. Kim, "Exploring author name disambiguation on PubMed-scale," *Journal of informetrics*, vol. 9, no. 4, pp. 924–941, 2015.

[45] B. Zhang, M. Dundar, and M. A. Hasan, "Bayesian Non-Exhaustive Classification for Active Online Name Disambiguation," *CoRR*, vol. abs/1708.0, 2017. [Online]. Available: http://arxiv.org/abs/1708.04531

[46] Z. Zhao, J. Rollins, L. Bai, and G. Rosen, "Incremental Author Name Disambiguation for Scientific Citation Data," in *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, oct 2017, pp. 175–183.

[47] M.-C. Müller, "On the contribution of word-level semantics to practical author name disambiguation," in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, 2018, pp. 367–368.

[48] B. Zhang, M. Dundar, V. Dave, and M. Hasan, "Dirichlet process gaussian mixture for active online name disambiguation by particle filter," in *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 2019, pp. 269–278.

[49] K. Kim, S. Rohatgi, and C. L. Giles, "Hybrid deep pairwise classification for author name disambiguation," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 2369–2372.

[50] D. Vishnyakova, R. Rodriguez-Esteban, and F. Rinaldi, "A new approach and gold standard toward author disambiguation in MEDLINE," *Journal of the American Medical Informatics Association*, vol. 26, no. 10, pp. 1037–1045, 2019.

[51] Q. Sun, H. Peng, J. Li, S. Wang, X. Dong, L. Zhao, S. Y. Philip, and L. He, "Pairwise learning for name disambiguation in large-scale heterogeneous academic networks," in *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2020, pp. 511–520.

[52] Z. Zhang, B. Yu, T. Liu, and D. Wang, "Strong Baselines for Author Name Disambiguation with and Without Neural Networks," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2020, pp. 369–381.

[53] Y. Chen, Z. Jiang, J. Gao, H. Du, L. Gao, and Z. Li, "A supervised and distributed framework for cold-start author disambiguation in large-scale publications," *Neural Computing and Applications*, pp. 1–16, 2021.

[54] Z. Boukhers, N. Bahubali, A. T. Chandrasekaran, A. Anand, S. M. G. Prasadand, and S. Aralappa, "Bib2Auth: Deep Learning Approach for Author Disambiguation using Bibliographic Data," *arXiv preprint arXiv:2107.04382*, 2021.

[55] Q. Li, "Co-attention-based pairwise learning for author name disambiguation," Master's thesis, University of Twente, 2022.

[56] Z. Boukhers and N. B. Asundi, "Deep author name disambiguation using dblp data," *International Journal on Digital Libraries*, pp. 1–11, 2023.

[57] C. L. Giles, H. Zha, and H. Han, "Name disambiguation in author citations using a K-way spectral clustering method," in *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 05)*, jun 2005, pp. 334–343.

[58] N. Aswani, K. Bontcheva, and H. Cunningham, "Mining Information for Instance Unification," in *The Semantic Web - ISWC 2006*, I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. M. Aroyo, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 329–342.

[59] A. P. de Carvalho, A. A. Ferreira, A. H. F. Laender, and M. A. Gonçalves, "Incremental unsupervised name disambiguation in cleaned digital libraries." *Journal of Information and Data Management*, vol. 2, pp. 289—-289, 2011.

[60] A. F. Santana, M. A. Gonçalves, A. H. F. Laender, and A. A. Ferreira, "Incremental Author Name Disambiguation by Exploiting Domain-specific Heuristics," *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 4, pp. 931–945, apr 2017. [Online]. Available: https://doi.org/10.1002/asi.23726

[61] J. Xu, S. Shen, D. Li, and Y. Fu, "A network-embedding based method for author disambiguation," in *Proceedings of the 27th ACM international conference on information and knowledge management*, 2018, pp. 1735–1738.

[62] S. Zhang, T. Huang, F. Yang, and Others, "ANDMC: An algorithm for author name disambiguation based on molecular cross clustering," in *International Conference on Database Systems for Advanced Applications*. Springer, 2019, pp. 173–185.

[63] J. Kim, J. Kim, and J. Owen-Smith, "Generating automatically labeled data for author name disambiguation: an iterative clustering method," *Scientometrics*, vol. 118, no. 1, pp. 253–280, 2019.

[64] B. Xiong, P. Bao, and Y. Wu, "Learning semantic and relationship joint embedding for author name disambiguation," *Neural Computing and Applications*, vol. 33, no. 6, pp. 1987–1998, 2021.

[65] X. Fan, J. Wang, X. Pu, L. Zhou, and B. Lv, "On graph-based name disambiguation," *Journal of Data and Information Quality (JDIQ)*, vol. 2, no. 2, pp. 1–23, 2011.

[66] B. Zhang and M. Al Hasan, "Name disambiguation in anonymized graphs using network embedding," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 1239–1248.

[67] I. Hussain and S. Asghar, "DISC: Disambiguating homonyms using graph structural clustering," *Journal of Information Science*, vol. 44, no. 6, pp. 830–847, 2018.

[68] W. Zhang, Z. Yan, and Y. Zheng, "Author name disambiguation using graph node embedding method," in *2019 IEEE 23rd international conference on computer supported cooperative work in design (CSCWD)*. IEEE, 2019, pp. 410–415.

[69] X. Ma, R. Wang, and Y. Zhang, "Author Name Disambiguation in Heterogeneous Academic Networks," in *International Conference on Web Information Systems and Applications*. Springer, 2019, pp. 126–137.

[70] Y. Ma, Y. Wu, and C. Lu, "A Graph-Based Author Name Disambiguation Method and Analysis via Information Theory," *Entropy*, vol. 22, no. 4, 2020. [Online]. Available: https://www.mdpi.com/1099-4300/22/4/416

[71] K. Pooja, S. Mondal, and J. Chandra, "Exploiting Higher Order Multi-dimensional Relationships with Self-attention for Author Name Disambiguation," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 16, no. 5, pp. 1–23, 2022.

[72] ——, "Online author name disambiguation in evolving digital library," *Neurocomputing*, vol. 493, pp. 1–14, 2022.

[73] P. Riehmann, M. Hanfler, and B. Froehlich, "Interactive sankey diagrams," in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* IEEE, 2005, pp. 233–240.

[74] S. Subramanian, D. King, D. Downey, and S. Feldman, "S2AND: A Benchmark and Evaluation System for Author Name Disambiguation," in *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL).* IEEE, 2021, pp. 170–179.

[75] D. Han, S. Liu, Y. Hu, B. Wang, and Y. Sun, "ELM-based name disambiguation in bibliography," *World Wide Web*, vol. 18, no. 2, pp. 253–263, mar 2015. [Online]. Available: https://doi.org/10.1007/s11280-013-0226-4

[76] X. Fan, J. Wang, X. Pu, L. Zhou, and B. Lv, "On Graph-Based Name Disambiguation," *J. Data and Information Quality*, vol. 2, no. 2, pp. 10:1—-10:23, feb 2011. [Online]. Available: http://doi.acm.org/10.1145/1891879.1891883

[77] L. V. B. Esperidião, A. A. Ferreira, A. H. F. Laender, M. A. Goncalves, D. Menotti, A. I. Tavares, and G. T. de Assis, "Reducing Fragmentation in Incremental Author Name Disambiguation," *JIDM*, vol. 5, pp. 293–307, 2014.

[78] H. N. Tran, T. Huynh, and T. Do, "Author name disambiguation by using deep neural network," in *Asian Conference on Intelligent Information and Database Systems.* Springer, 2014, pp. 123–132.

[79] H. Hazimeh, I. Youness, J. Makki, H. Noureddine, J. Tscherrig, E. Mugellini, and O. A. Khaled, "Leveraging Co-authorship and Biographical Information for Author Ambiguity Resolution in DBLP," *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, pp. 1080–1084, 2016.

[80] M.-C. Müller, "Semantic author name disambiguation with word embeddings," in *International conference on theory and practice of Digital Libraries.* Springer, 2017, pp. 300–311.

[81] S. Mishra, S. Saha, and S. Mondal, "GAEMTBD: Genetic algorithm based entity matching techniques for bibliographic databases," *Applied Intelligence*, vol. 47, no. 1, pp. 197–230, jul 2017. [Online]. Available: https://doi.org/10.1007/s10489-016-0874-z

[82] N. Kooli, R. Allesiardo, and E. Pigneul, "Deep learning based approach for entity resolution in databases," in *Asian Conference on Intelligent Information and Database Systems*. Springer, 2018, pp. 3–12.

[83] J. Kim, "A fast and integrative algorithm for clustering performance evaluation in author name disambiguation," *Scientometrics*, vol. 120, no. 2, pp. 661–681, 2019.

[84] S. Zhang, E. Xinhua, and T. Pan, "A multi-level author name disambiguation algorithm," *IEEE Access*, vol. 7, pp. 104 250–104 257, 2019.

[85] Z. Boukhers and N. B. Asundi, "Deep author name disambiguation using bibliographic data," in *International Conference on Theory and Practice of Digital Libraries*. Springer, 2022, pp. 201–215.

[86] V. I. Torvik and N. R. Smalheiser, "Author name disambiguation in MEDLINE," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, no. 3, p. 11, 2009.

[87] J. Kim and J. Owen-Smith, "ORCID-linked labeled data for evaluating author name disambiguation at scale," *Scientometrics*, vol. 126, no. 3, pp. 2057–2083, 2021.

[88] L. Zhang, W. Lu, and J. Yang, "LAGOS-AND: A Large, Gold Standard Dataset for Scholarly Author Name Disambiguation," *arXiv preprint arXiv:2104.01821*, 2021.

[89] V. I. Torvik and S. Agarwal, "Ethnea–an instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database," 2016.

[90] C. Santini, G. A. Gesese, S. Peroni, A. Gangemi, H. Sack, and M. Alam, "A knowledge graph embeddings based approach for author name disambiguation using literals," *Scientometrics*, vol. 127, no. 8, pp. 4887–4912, 2022.

[91] J. Heaton, "Ian goodfellow, yoshua bengio, and aaron courville: Deep learning: The mit press, 2016, 800 pp, isbn: 0262035618," *Genetic programming and evolvable machines*, vol. 19, no. 1-2, pp. 305–307, 2018.

[92] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[93] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[94] E. Achtert, S. Goldhofer, H.-P. Kriegel, E. Schubert, and A. Zimek, "Evaluation of clusterings–metrics and visual support," in *2012 IEEE 28th International Conference on Data Engineering*. IEEE, 2012, pp. 1285–1288.

[95] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica: Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[96] F. Wikipedia, "Google scholar," 2020. [Online]. Available: https://en.wikipedia.org/wiki/Google_Scholar

[97] D. G. Altman and J. M. Bland, "Statistics Notes: Detecting skewness from summary information," *Bmj*, vol. 313, no. 7066, p. 1200, 1996.

Appendix A

Tables

Table A.1: Intermediate Feature Ranking Based on the Usefulness of a Feature.

Features	**f**	**d**	**u**	**Rank**	**Rank#**
Author Name Variants	9	9	1.00	9.00	1
Co-authors name	9	5	0.56	5.00	2
Paper title	8	4	0.44	3.56	3
Author email	5	5	0.56	2.78	4
Author affiliation	5	4	0.44	2.22	5
Key phrases/Keywords	3	3	0.33	1.00	6
Paper abstract	3	2	0.22	0.67	7
Journal year	3	2	0.22	0.67	7
Organization	3	2	0.22	0.67	7
Publication venue/Journal shared	5	1	0.11	0.56	8
Mesh shared	2	2	0.22	0.44	9
Author research area/Major	2	1	0.11	0.22	10
Location	2	1	0.11	0.22	10
Languages	2	1	0.11	0.22	10
Address	1	1	0.11	0.11	11

TABLE A.2: Impact of Feature (list1) Combinations, using Arnetminer.

	Feature(s)	**pP**	**pR**	**pF1**	**ACP**	**AAP**	**K**	**CP**	**CR**	**CF1**	**PC**	**AC**	**Δ pF1**	**Δ K**
A	Coauthors	1.00	0.74	0.85	1.00	0.68	0.83	0.46	0.76	0.57	1370	825		
B	A+Paper Title	1.00	0.74	0.85	1.00	0.68	0.83	0.46	0.76	0.57	1370	825	0%	0%
C	**B+Affiliation**	0.99	0.91	0.95	1.00	0.83	0.91	0.62	0.82	0.71	1092	825	**11%**	**10%**
D	C+Year	0.97	0.91	0.94	0.99	0.83	0.91	0.63	0.82	0.71	1084	825	-1%	-1%
E	D+Venue	0.89	0.94	0.91	0.90	0.88	0.89	0.77	0.85	0.81	908	825	-3%	-2%

TABLE A.3: Impact of Feature (list1) Combinations, using CustAND.

	Feature(s)	**pP**	**pR**	**pF1**	**ACP**	**AAP**	**K**	**CP**	**CR**	**CF1**	**PC**	**AC**	**Δ pF1**	**Δ K**
A	Coauthors	0.95	0.52	0.67	0.98	0.51	0.71	0.11	0.34	0.17	320	105		
B	A+Paper Title	0.95	0.52	0.67	0.98	0.51	0.71	0.11	0.34	0.17	320	105	0%	0%
C	**B+Email**	0.97	0.96	0.96	0.97	0.92	0.94	0.56	0.71	0.63	135	105	**43%**	**33%**
D	**C+Affiliation**	0.97	0.99	0.98	0.97	0.96	0.96	0.73	0.83	0.78	119	105	**2%**	**2%**
E	D+Keywords	0.97	0.99	0.98	0.97	0.96	0.96	0.73	0.83	0.78	119	105	0%	0%
F	E+Abstract	0.97	0.99	0.98	0.97	0.96	0.96	0.73	0.83	0.78	119	105	0%	0%
F'	E+Year	0.97	0.99	0.98	0.97	0.96	0.96	0.73	0.83	0.78	119	105	0%	0%
G	F+Venue	0.93	0.99	0.96	0.94	0.97	0.96	0.82	0.87	0.84	111	105	-1%	-1%
G'	F'+Venue	0.93	0.99	0.96	0.94	0.97	0.96	0.82	0.87	0.84	111	105	-1%	-1%

Table A.4: Impact of Feature (list1) Combinations, using PubMed.

	Feature(s)	**pP**	**pR**	**pF1**	**ACP**	**AAP**	**K**	**CP**	**CR**	**CF1**	**PC**	**AC**	**Δ pF1**	**Δ K**
A	Coauthors	0.84	0.31	0.45	0.91	0.62	0.75	0.35	0.64	0.45	1303	703		
	Name													
B	A+Paper Title	0.84	0.31	0.45	0.91	0.62	0.75	0.35	0.64	0.45	1303	703	0%	0%
C	**B+Email**	0.83	0.73	0.78	0.87	0.77	0.82	0.55	0.72	0.62	933	703	**71%**	**9%**
D	C+Affiliation	0.69	0.87	0.77	0.79	0.87	0.83	0.88	0.82	0.85	650	703	-1%	1%
E	D+Abstract	0.69	0.87	0.77	0.79	0.87	0.83	0.88	0.82	0.85	650	703	0%	0%
E'	D+Year	0.64	0.87	0.74	0.78	0.87	0.83	0.89	0.82	0.85	646	703	-4%	0%
G	E+Venue	0.44	0.95	0.60	0.66	0.92	0.78	0.86	0.86	0.86	508	703	-21%	-6%
G'	E'+Venue	0.44	0.95	0.60	0.66	0.92	0.78	0.86	0.86	0.86	508	703	-18%	-6%

Table A.5: Impact of Feature (list2) Combinations, using Arnetminer.

	Feature(s)	**pP**	**pR**	**pF1**	**ACP**	**AAP**	**K**	**CP**	**CR**	**CF1**	**PC**	**AC**	**Δ pF1**	**Δ K**
A	Coauthor	1	0.74	0.85	1	0.69	0.83	0.46	0.76	0.58	1394	845		
	Name													
B	**A+Venue**	0.91	0.88	0.9	0.92	0.81	0.86	0.63	0.8	0.7	1073	845	**5%**	**3%**
C	B+Paper Title	0.91	0.88	0.9	0.92	0.81	0.86	0.63	0.8	0.7	1073	845	0%	0%
D	**C+Affiliation**	0.89	0.94	0.91	0.9	0.88	0.89	0.77	0.85	0.81	931	845	**2%**	**4%**
E	D+Year	0.89	0.94	0.91	0.9	0.88	0.89	0.77	0.85	0.81	931	845	0%	0%

TABLE A.6: Impact of Feature (list2) Combinations, using CustAND.

	Feature(s)	pP	pR	pF1	ACP	AAP	K	CP	CR	CF1	PC	AC	Δ pF1	Δ K
A	Coauthor Name	0.95	0.52	0.67	0.98	0.51	0.71	0.11	0.34	0.17	320	105		
B	**A+Venue**	0.94	0.73	0.82	0.97	0.65	0.79	0.17	0.39	0.23	246	105	**22%**	**12%**
C	B+Title	0.94	0.73	0.82	0.97	0.65	0.79	0.17	0.39	0.23	246	105	0%	0%
D	**C+Affiliation**	0.94	0.96	0.95	0.95	0.91	0.93	0.54	0.7	0.61	136	105	**16%**	**18%**
E	D+Abstract	0.94	0.96	0.95	0.95	0.91	0.93	0.54	0.7	0.61	136	105	0%	0%
F	E+Keywords	0.94	0.96	0.95	0.95	0.91	0.93	0.54	0.7	0.61	136	105	0%	0%
G	F+Year	0.94	0.96	0.95	0.95	0.91	0.93	0.54	0.7	0.61	136	105	0%	0%
H	**G+Email**	0.93	0.99	0.96	0.94	0.97	0.96	0.82	0.87	0.84	111	105	**1%**	**3%**

TABLE A.7: Impact of Feature (list2) Combinations, using PubMed.

	Feature(s)	pP	pR	pF1	ACP	AAP	K	CP	CR	CF1	PC	AC	Δ pF1	Δ K
A	Coauthor Name	0.8	0.3	0.45	0.91	0.62	0.75	0.3	0.6	0.4	1303	703		
B	**A+Venue**	0.5	0.6	0.55	0.75	0.76	0.76	0.6	0.7	0.7	836	703	**22%**	**1%**
C	B+Paper Title	0.5	0.6	0.55	0.75	0.76	0.76	0.6	0.7	0.7	836	703	0%	0%
D	**C+Affiliation**	0.4	0.9	0.61	0.67	0.9	0.77	1	0.8	0.9	552	703	**10%**	**2%**
E	D+Abstract	0.4	0.9	0.61	0.67	0.9	0.77	1	0.8	0.9	552	703	0%	0%
F	E+Year	0.4	0.9	0.61	0.67	0.9	0.77	1	0.8	0.9	552	703	0%	0%
G	F+Email	0.4	0.9	0.6	0.66	0.92	0.78	0.9	0.9	0.9	508	703	-1%	0%

TABLE A.8: (a) Impact of Feature (list3) Combinations on Authorship Results, using Arnetminer.

	Arnetminer						
	Features/combinations	**pP**	**pR**	**pF1**	**ACP**	**AAP**	**K**
A	Co-authors Names	1	0.74	0.85	1	0.63	0.8
B	**Author Affiliation**	0.97	0.76	**0.86**	0.98	0.6	0.77
C	Author Email	NA	NA	NA	NA	NA	NA
D	Paper Venue	0.957	0.083	0.15	0.972	0.257	0.5
E	Paper Title	0.603	0.915	0.727	0.593	0.878	0.721
A'	**A+B**	0.99	0.907	**0.947**	1	0.799	0.895
D'	A+D	0.927	0.88	0.903	0.933	0.77	0.847
E'	A+E	0.901	0.688	0.78	0.95	0.593	0.75
D"	B+D	0.912	0.887	0.9	0.91	0.723	0.811
E"	B+E	0.94	0.79	0.86	0.95	0.66	0.79
A"	A+B+D	0.915	0.937	0.926	0.921	0.856	0.888
A"'	**A+B+E**	0.983	0.92	**0.95**	0.99	0.82	0.9

TABLE A.9: (b) Impact of Feature (list3) Combinations on Authorship Results, using CustAND.

	CustAND						
	Features/combinations	**pP**	**pR**	**pF1**	**ACP**	**AAP**	**K**
A	Co-authors Names	1	0.476	0.645	0.999	0.442	0.665
B	Author Affiliation	0.977	0.72	0.829	0.977	0.731	0.845
C	**Author Email**	0.949	0.856	**0.9**	0.949	0.801	0.872
D	Paper Venue	0.916	0.098	0.17	0.958	0.209	0.447
E	Paper Title	0.591	0.656	0.622	0.745	0.61	0.674
A'	C+A	0.95	0.87	0.91	0.948	0.825	0.885
B'	**C+B**	1	0.96	**0.98**	0.99	0.965	0.975
D'	C+D	0.95	0.877	0.912	0.949	0.825	0.885
E'	C+E	0.849	0.914	0.88	0.902	0.851	0.876
A"	C+B+A	0.964	0.92	0.95	0.948	0.875	0.91
C"	C+B+D	0.951	0.9	0.925	0.938	0.87	0.90
E"	C+B+E	0.851	0.95	0.90	0.901	0.91	0.91
EE	C+B+A+E	0.94	0.94	0.94	0.95	0.896	0.922
EE'	C+B+D+E	0.93	0.91	0.92	0.938	0.886	0.91

TABLE A.10: (c) Impact of Feature (list3) Combinations on Authorship Results, using PubMed.

		PubMed					
	Features/combinations	**pP**	**pR**	**pF1**	**ACP**	**AAP**	**K**
A	Co-authors Name	0.84	0.31	0.45	0.91	0.62	0.75
B	**Author Affiliation**	0.9	0.63	**0.74**	0.9	0.62	0.75
C	Author Email	0.97	0.42	0.59	0.97	0.43	0.65
D	Paper Venue	0.74	0.11	0.19	0.9	0.37	0.58
E	Paper Title	1	0.01	0.02	1	0.25	0.5
F	B+A	0.64	0.9	0.75	0.79	0.85	0.82
F1	**B+C**	0.89	0.69	**0.77**	0.88	0.7	0.79
F2	B+D	0.5	0.83	0.62	0.74	0.76	0.75
F3	B+E	0.9	0.63	0.74	0.9	0.62	0.75
G1	B+A+C	0.69	0.87	0.77	0.79	0.87	0.83
G2	B+A+D	0.45	0.94	0.61	0.67	0.9	0.77
G3	B+A+E	0.7	0.85	0.77	0.8	0.83	0.82
G4	B+C+A	0.62	0.92	0.74	0.77	0.89	0.83
G5	B+C+D	0.48	0.85	0.61	0.72	0.81	0.76
G6	B+C+E	0.89	0.69	0.77	0.88	0.7	0.79
H1	B+A+C+D	0.43	0.95	0.59	0.65	0.92	0.77
H2	B+A+C+E	0.62	0.91	0.74	0.77	0.89	0.83

TABLE A.11: Individual Feature Rankings Based on pF1 Ccores.

Dataset	**Feature**	**pP**	**pR**	**pF1**	**Rank#**
Arnetminer	author affiliation	97%	76%	86%	1
	co-authors	100%	74%	85%	2
	paper title	60%	91%	73%	3
	paper venue	96%	10%	15%	4
CustAND	author email	95%	86%	90%	1
	author affiliation	97%	72%	83%	2
	co-authors	100%	48%	65%	3
	paper title	60%	66%	62%	4
	paper venue	92%	10%	17%	5
PubMed	author affiliation	90%	63%	74%	1
	author email	97%	42%	59%	2

co-authors	84%	31%	45%	3
paper venue	74%	11%	19%	4
paper title	100%	1%	1%	5

TABLE A.12: Statistics of the Sample Data used.

Variable	**Obs**	**Missing data**	**Min value**	**Max value**	**Mean**	**Std. deviation**
Co-authors Names	501	0	0	0.7	0.085	0.132
Paper Title	501	0	0	0.8	0.054	0.095
Paper Venue	501	0	0	0.4	0.033	0.075
Paper Affiliation	501	0	0	0.9	0.443	0.303
Paper Email	501	0	0	0.9	0.222	0.304

TABLE A.13: Eigen values of the Factors with Respect to the Individual and Cumulative Variability.

	F1	**F2**	**F3**	**F4**	**F5**
Eigenvalue	1.682	1.453	0.91	0.719	0.236
Variability (%)	33.645	29.063	18.193	14.371	4.728
Cumulative %	33.645	62.708	80.901	95.272	100

TABLE A.14: Correlations Between Features and Factors.

Feature Name	**F1**	**F2**	**F3**	**F4**	**F5**
Co-authors Names	-0.431	-0.534	-0.585	-0.409	0.141
Paper Title	-0.43	-0.628	0.07	0.635	0.11
Paper Venue	-0.056	-0.586	0.714	-0.378	0.025
Paper Affiliation	-0.66	0.646	0.227	-0.055	0.305
Paper Email	0.935	-0.114	-0.035	0.042	0.333

TABLE A.15: Kappa Calculation (Rater 1 versus Rater 2).

		Annotator 1		**Row marginals**	
		Normal	Abnormal		
Annotator 2	Normal	7886	20	**7906**	rm^1
	Abnormal	10	435	**460**	rm^2

Column Marginals	**7896**	**470**	**8366**	n
	cm^1	cm^2	n	

$$Pr_{(a)} = \frac{(7886 + 450)}{8366} = 1$$

$$Pr_{(e)} = \frac{\frac{7896 * 7906}{8366} + \frac{470 * 460}{8366}}{8366} = \frac{(7461.84 + 25.84)}{8366} = 0.895$$

$$\kappa = \frac{(1 - 0.895)}{(1 - 0.895)} = \frac{0.105}{0.105} = 1 \tag{A.1}$$

TABLE A.16: Kappa Calculation (Rater 1 versus Rater 3).

		Annotator 1		**Row marginals**	
		Normal	Abnormal		
Annotator 3	Normal	7886	15	**7901**	rm^1
	Abnormal	20	435	**455**	rm^2
Column Marginals		**7906**	**450**	**8356**	n
		cm^1	cm^2	n	

$$Pr_{(a)} = \frac{(7886 + 435)}{8356} = 0.9958$$

$$Pr_{(e)} = \frac{\frac{7906 * 7901}{8366} + \frac{450 * 455}{8356}}{8356} = \frac{(7475.5 + 24.5)}{8356} = 0.897$$

$$\kappa = \frac{(0.99 - 0.897)}{(1 - 0.897)} = \frac{0.098}{0.103} = 0.96 \tag{A.2}$$

TABLE A.17: Kappa Calculation (Rater 2 versus Rater 3).

		Annotator 2		**Row marginals**	
		Normal	Abnormal		
Annotator 3	Normal	7886	25	**7911**	rm^1
	Abnormal	10	440	**450**	rm^2

Column Marginals	**7896**	**465**	**8361**	n
	cm^1	cm^2	n	

$$Pr_{(a)} = \frac{(7886 + 440)}{8361} = 0.9958$$

$$Pr_{(e)} = \frac{\frac{7896 * 7911}{8361} + \frac{465 * 450}{8361}}{8361} = \frac{(7471 + 25.02)}{8361} = 0.896$$

$$\kappa = \frac{(0.9958 - 0.896)}{(1 - 0.896)} = \frac{0.0998}{0.104} = 0.96 \tag{A.3}$$

Table A.18: The BDBComp Dataset Details.

Ambiguous group	**#of references/ # of authors**	**Ambiguous group**	**#of references/ # of authors**
A. Oliveira	52/20	J. Souza	34/12
A. Silva	64/38	L. Silva	33/18
F. Silva	27/22	M. Silva	21/16
J. Oliveira	48/22	R. Santos	20/17
J. Silva	35/18	R. Silva	27/22

Table A.19: The Arnetminer Dataset Details.

Group	**Ref/Au**	**Group**	**Ref/Au**	**Group**	**Ref/Au**	**Group**	**Ref/Au**	**Group**	**Ref/Au**
Ajay Gupta	36/9	Frank Mueller	101/3	Keith Edwards	23/4	R. Ramesh	46/9	Yang Yu	71/19
Alok Gupta	57/2	Gang Chen	178/47	Koichi Furukawa	77/2	Rafael Alonso	40/2	Yi Deng	89/ 9
Barry Wilkinson	28/1	Gang Luo	47/9	Kuo Zhang	16/4	Rakesh Kumar	96/10	Yong Chen	84/25
Bin Li	181/60	Hao Wang	178/48	Lei Chen	196/40	Richard Taylor	35/16	Yoshio Tanaka	43/2
Bin Yu	105/17	Hiroshi Tanaka	40/7	Lei Fang	17/7	Robert Allen	24/9	Young Park	21/9
Bin Zhu	46/15	Hong Xie	12/7	Lei Jin	16/6	Robert Schreiber	58/1	Yu Zhang	235/72
Bing Liu	182/18	Hui Fang	42/8	Lei Wang	308/112	S. Huang	16/15	Yue Zhao	41/9

Bo Liu	124/47	Hui Yu	32/21	Li Shen	68/9	Sanjay Jain	217/5	Yun Wang	46/19
Bob Johnson	11/7	J. Guo	13/10	Lu Liu	58/17	Satoshi Kobayashi	38/4	Z. Wang	47/38
Charles Smith	7/4	J. Yin	18/7	M. Rahman	17/9	Shu Lin	76/2	Xiaoyan Li	33/6
Cheng Chang	27/5	Jeffrey Parsons	31/2	Manuel Silva	74/4	Steve King	31/4	Yan Tang	32/11
Daniel Massey	43/2	Ji Zhang	64/16	Mark Davis	24/6	Thomas D. Taylor	4/3	Yang Wang	195/55
David Brown	61/25	Jianping Wang	37/5	Michael Lang	17/4	Thomas Hermann	44/8	Qiang Shen	70/3
David C. Wilson	65/5	Jie Tang	66/6	Michael Siegel	54/6	Thomas Meyer	31/7	R. Balasubramanian	20/6
David Cooper	18/7	Jie Yu	32/9	Michael Smith	33/19	Thomas Tran	15/2	R. Cole	22/5
David E. Goldberg	231/3	Jim Gray	192/6	Michael Wagner	71/14	Thomas Wolf	33/8	Kai Tang	48/3
David Jensen	53/4	Jing Zhang	231/85	Ning Zhang	127/33	Thomas Zimmermann	67/2	Kai Zhang	66/24
David Levine	48/18	John Collins	27/7	Paul Brown	27/8	Wei Xu	153/48	Ke Chen	107/16
David Nelson	20/11	John F. McDonald	34/2	Paul Wang	16/7	Wen Gao	484/10	Fei Su	37/4
Éric Martin	85/5	John Hale	39/4	Peter Phillips	13/3	William H. Hsu	34/2	Feng Liu	149/32
F. Wang	19 /17	Jose M. García	83/2	Philip J. Smith	33/3	X. Zhang	62/40	Feng Pan	73/15
Fan Wang	56/14	Juan Carlos Lopez	36/1	Ping Zhou	36/18	Xiaoming Wang	38/14		

Ref = no of references, Au = no of authors

TABLE A.20: Result Comparison of MHCF with MHCF-G and MHCF-GL using Complete Arnetminer and BDBComp Datasets. Where the features used for Arnetminer dataset are(Author affiliation, co-authors, paper title, paper venue) and for BDBComp (Co-authors, paper title, paper venue)

Arnetminer											
Technique	**pP**	**pR**	**pF1**	**ACP**	**AAP**	**K**	**cP**	**cR**	**cF1**	**Predicted clusters**	**Actual clusters**
MHCF	**0.85**	0.9	**0.88**	0.88	0.85	0.87	0.71	0.83	0.77	1811	1546
MHCF-G	0.77	0.91	0.84	0.83	0.86	0.85	0.75	0.84	0.79	1718	1546
MHCF-GL	0.71	0.92	0.8	0.78	0.88	0.83	0.82	0.84	0.83	1595	1546
BDBComp											
MHCF	**0.88**	0.86	**0.86**	0.94	0.9	0.92	0.85	0.91	0.88	218	205
MHCF-G	0.82	0.88	0.83	0.9	0.91	0.9	0.88	0.91	0.89	211	205
MHCF-GL	0.87	0.87	0.85	0.93	0.9	0.91	0.85	0.9	0.88	217	205

TABLE A.21: Results Comparison of MHCF(M)(Co-authors, paper titles, venue) with SAND1(S1)(Author names, affiliation, publication venue) and SAND2(S2)(Author names, affiliation, publication venue) using BDBComp Dataset.

	S1		S2		M		M vs S1		M vs S2	
Ambiguous Name	K	pF1	K	pF1	K	pF1	Δ K	Δ pF1	Δ K	Δ pF1

a oliveira	0.84	0.71	0.93	0.903	0.93	0.79	**11%**	**11%**	0%	-13%
a silva	0.95	0.835	0.982	0.971	0.98	0.87	**3%**	**4%**	0%	-10%
f silva	0.95	0.71	0.95	0.71	0.95	0.98	0%	**38%**	0%	**38%**
j oliveira	0.92	0.87	0.83	0.68	0.83	0.82	-10%	-6%	0%	**21%**
j silva	0.91	0.72	0.95	0.93	0.95	0.81	**4%**	**13%**	0%	-13%
j souza	0.75	0.44	0.94	0.904	0.94	0.78	**25%**	**77%**	0%	-14%
l silva	0.84	0.6	0.9	0.737	0.9	0.91	**7%**	**52%**	0%	**23%**
m silva	0.93	0.57	0.95	0.74	0.95	0.88	**2%**	**54%**	0%	**19%**
r santos	0.98	0.8	0.91	0.48	0.91	0.91	-7%	**14%**	0%	**90%**
r silva	0.9	0.55	0.9	0.47	0.9	0.84	0%	**53%**	0%	**79%**
	90%	68%	92%	75%	92%	**86%**	**4%**	**31%**	0%	**22%**

TABLE A.22: Result Comparison of MHCF(Co-authors, paper title, paper venue) with HHC(using all features) using BDBComp Dataset.

Technique	**ACP**	**AAP**	**K**	**pP**	**pR**	**pF1**	**Δ pF1**	**Δ K**
HHC	0.88	0.99	0.93	0.58	0.83	0.65		
MHCF	0.94	0.9	**0.92**	0.88	0.86	**0.86**	**32%**	-1%

TABLE A.23: MHCF Overall Results in Comparison with GFAD-AD and GFAD-OR using two features (Co-authors and paper title)

	MHCF		**GFAD-AD**		**GFAD-OR**		**MHCF v GFAD-AD**		**MHCF v GFAD-OR**	
Ambiguous Name	**pF1**	**K**	**pF1**	**K**	**pF1**	**K**	**Δ pF1**	**Δ K**	**Δ pF1**	**Δ K**
A. Gupta	0.53	0.63	0.55	0.57	0.52	0.55	-5%	10%	1%	14%
Bin Li	0.88	0.92	0.4	0.57	0.37	0.56	120%	61%	138%	64%
Charles Smith	1	1	0.97	1	0.97	1	3%	0%	3%	0%
Daniel Massey	0.79	0.82	0.89	0.91	0.89	0.91	-11%	-10%	-11%	-10%
David C. Wilson	0.57	0.67	0.62	0.7	0.62	0.69	-8%	-4%	-8%	-3%
David E. Goldberg	0.77	0.8	0.96	0.96	0.96	0.96	-20%	-17%	-20%	-17%
Éric Martin	0.91	0.89	0.71	0.73	0.71	0.73	28%	22%	28%	22%
Fei Su	0.95	0.95	0.88	0.85	0.91	0.87	8%	12%	4%	9%
Jie Yu	0.99	0.94	0.97	0.94	0.97	0.94	2%	0%	2%	0%
John Collins	0.95	0.96	0.43	0.49	0.43	0.49	121%	96%	121%	96%
John Hale	0.65	0.71	0.7	0.71	0.68	0.71	-7%	0%	-4%	0%
B. Liu	0.56	0.76	0.63	0.68	0.6	0.67	-11%	12%	-7%	13%
David Brown	0.99	0.98	0.97	0.89	0.97	0.87	2%	10%	2%	13%
David Cooper	0.95	0.96	0.97	0.77	0.97	0.77	-2%	25%	-2%	25%

Gang Luo	0.83	0.9	0.97	0.98	0.97	0.98	-14%	-8%	-14%	-8%
Hiroshi Tanaka	0.47	0.69	0.51	0.6	0.51	0.6	-8%	15%	-8%	15%
Jeffrey Parsons	0.25	0.44	0.28	0.44	0.28	0.44	-11%	0%	-11%	0%
Jie Tang	0.91	0.93	0.94	0.96	0.94	0.96	-3%	-3%	-3%	-3%
John F. McDonald	0.94	0.94	0.96	0.97	0.96	0.97	-2%	-3%	-2%	-3%
Jose M. García	0.71	0.76	0.86	0.87	0.86	0.87	-17%	-13%	-17%	-13%
Manuel Silva	0.78	0.83	0.83	0.84	0.83	0.84	-6%	-1%	-6%	-1%
Michael Lang	0.65	0.82	0.59	0.5	0.59	0.5	10%	64%	10%	64%
S. Huang	1	1	0.96	0.92	0.96	0.92	4%	9%	4%	9%
Wen Gao	0.95	0.95	0.29	0.48	0.33	0.51	228%	98%	188%	86%
X. Zhang	0.92	0.95	0.67	0.67	0.67	0.67	37%	42%	37%	42%
Xiaoyan Li	0.87	0.92	0.97	0.75	0.97	0.72	-10%	23%	-10%	28%
Yan Tang	0.9	0.92	0.97	0.87	0.97	0.84	-7%	6%	-7%	10%
Yi Deng	0.93	0.93	0.97	0.79	0.97	0.76	-4%	18%	-4%	22%
Lei Jin	1	1	0.97	0.8	0.97	0.8	3%	25%	3%	25%
Li Shen	0.83	0.85	0.73	0.77	0.73	0.77	14%	10%	14%	10%
Michael Siegel	0.87	0.87	0.85	0.83	0.85	0.83	2%	5%	2%	5%
Michael Wagner	0.51	0.74	0.68	0.68	0.67	0.67	-25%	9%	-24%	10%

R. Ramesh	0.66	0.74	0.34	0.51	0.34	0.51	94%	45%	94%	45%
Rakesh Kumar	0.89	0.86	0.97	0.73	0.97	0.7	-8%	18%	-8%	23%
Robert Allen	0.8	0.89	0.97	1	0.97	1	-18%	-11%	-18%	-11%
Sanjay Jain	0.79	0.78	0.44	0.53	0.51	0.58	80%	47%	55%	34%
Shu Lin	0.8	0.82	0.35	0.43	0.35	0.44	129%	91%	129%	86%
Thomas D. Taylor	1	1	0.62	0.65	0.62	0.65	61%	54%	61%	54%
Yue Zhao	0.69	0.8	0.97	0.9	0.97	0.89	-29%	-11%	-29%	-10%
Average	**0.81**	**0.85**	0.75	0.75	0.75	0.75	**18%**	**19%**	**18%**	**19%**

TABLE A.24: MHCF Results of 11 Ambiguous Author Names used by MDC.

Ambiguous Names	**MHCF with 2 features (co-authors and paper title)**		**MHCF with 3 features (co-authors, author affiliation, paper title)**	
	pF1	**K**	**pF1**	**K**
Bin Li	0.88	0.92	0.94	0.95
Bo Liu	0.71	0.91	0.78	0.92
Feng Liu	0.56	0.75	0.59	0.77
Gang Chen	0.65	0.83	0.59	0.8

Lei Wang	0.77	0.88	0.69	0.87
Ning Zhang	0.69	0.82	0.78	0.87
Wei Xu	0.79	0.85	0.89	0.89
X. Zhang	0.92	0.95	0.92	0.95
Yang Wang	0.46	0.79	0.61	0.84
Yu Zhang	0.65	0.81	0.74	0.83
Z. Wang	0.92	0.95	0.92	0.95
Average	**0.73**	**0.86**	**0.77**	**0.87**

TABLE A.25: MHCF(Author affiliation, co-authors, paper title, paper venue) Comparison with ESMD(Co-authors, paper title, abstract, venues, references, affiliation) and ATGEP(Coauthors, abstract, reference, keywords from reference title, author profile information from external source).

	ESMD			**MHCF**			**ATGEP**		
Ambigious Names	**pP**	**pR**	**pF1**	**pP**	**pR**	**pF1**	**pP**	**pR**	**pF1**
Lu Liu	1.00	1.00	1.00	0.96	1.00	0.98	0.47	0.63	0.54
Mark Davis	1.00	1.00	1.00	1.00	0.99	0.99	0.95	0.57	0.72
John Hale	1.00	1.00	1.00	1.00	0.69	0.82	0.99	0.37	0.54
Kuo Zhang	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.85	0.92
Yue Zhao	1.00	1.00	1.00	1.00	1.00	1.00	0.92	0.98	0.95

Wen Gao	0.98	1.00	0.99	0.95	0.99	0.99	0.98	0.94	0.96
David Levine	0.69	0.55	0.61	1.00	1.00	1.00	0.68	0.52	0.59
Manuel Silva	0.79	0.91	0.85	0.97	0.94	0.97	0.81	0.87	0.84
Kai Zhang	0.81	0.87	0.84	1.00	0.98	0.99	0.14	0.86	0.24
Shu lin	1.00	1.00	1.00	1.00	0.80	0.89	1.00	0.90	0.95
Keith Edwards	1.00	1.00	1.00	1.00	0.77	0.87	0.85	0.36	0.51
Sanjay Jain	0.79	0.99	0.88	1.00	0.98	0.99	1.00	0.91	0.95
Bing Liu	0.80	0.98	0.88	0.97	0.78	0.86	0.85	0.68	0.76
R Ramesh	0.53	0.97	0.69	1.00	1.00	1.00	0.90	0.48	0.63
Average	0.89	0.95	0.91	**0.99**	0.91	**0.95**	0.84	0.69	0.71

TABLE A.26: MHCF Results on 109 Ambiguous Authors using Four Features (Co-authors, Author Affiliation, Paper Title, and Paper Venue).

Ambiguous Names	**pP**	**pR**	**pF1**	**ACP**	**AAP**	**K**
Ajay Gupta	1	0.64	0.78	1	0.64	0.8
Alok Gupta	1	0.88	0.93	1	0.85	0.92
Barry Wilkinson	1	0.38	0.55	1	0.4	0.63
Bin Li	0.54	0.93	0.68	0.72	0.93	0.82
Bin Yu	0.74	0.48	0.58	0.87	0.68	0.77
Bin Zhu	0.57	0.74	0.64	0.76	0.86	0.81
Bing Liu	0.81	0.81	0.81	0.86	0.83	0.85

Bo Liu	0.28	0.99	0.43	0.58	0.99	0.75
Bob Johnson	1	0.78	0.88	1	0.88	0.94
Charles Smith	1	1	1	1	1	1
Cheng Chang	1	0.68	0.81	1	0.77	0.88
Daniel Massey	1	0.78	0.87	1	0.79	0.89
David Brown	1	0.98	0.99	0.98	0.96	0.97
David C. Wilson	1	0.93	0.96	0.98	0.93	0.95
David Cooper	1	0.91	0.95	1	0.92	0.96
David E. Goldberg	1	0.99	1	1	0.99	1
David Jensen	0.96	0.8	0.87	0.96	0.8	0.88
David Levine	0.94	0.99	0.97	0.92	0.95	0.93
David Nelson	0.91	0.84	0.87	0.93	0.92	0.93
Éric Martin	1	0.91	0.95	1	0.88	0.94
F. Wang	0.89	1	0.94	0.93	1	0.96
Fan Wang	0.97	0.95	0.96	0.95	0.9	0.93
Fei Su	1	0.93	0.96	1	0.95	0.97
Feng Liu	0.58	0.45	0.5	0.76	0.62	0.69
Feng Pan	0.9	0.57	0.7	0.92	0.71	0.81
Frank Mueller	1	0.75	0.86	1	0.76	0.87
Gang Chen	0.49	0.77	0.6	0.67	0.84	0.75
Gang Luo	1	0.87	0.93	1	0.92	0.96

Hao Wang	0.72	0.7	0.71	0.82	0.78	0.8
Hiroshi Tanaka	0.98	0.38	0.55	0.96	0.53	0.71
Hong Xie	0.83	0.56	0.67	0.92	0.72	0.81
Hui Fang	0.98	0.98	0.98	0.96	0.96	0.96
Hui Yu	0.85	0.97	0.91	0.91	0.97	0.94
J. Guo	0.75	0.9	0.82	0.85	0.92	0.88
J. Yin	0.81	0.8	0.8	0.9	0.85	0.87
Jeffrey Parsons	1	0.51	0.68	1	0.54	0.74
Ji Zhang	0.78	0.94	0.85	0.89	0.91	0.9
Jianping Wang	1	0.94	0.97	1	0.95	0.97
Jie Tang	1	1	1	1	1	1
Jie Yu	1	0.99	1	1	0.97	0.98
Jim Gray	0.99	0.79	0.88	0.98	0.79	0.88
Jing Zhang	0.09	0.78	0.17	0.44	0.88	0.62
John Collins	1	1	1	1	1	1
John F. McDonald	1	0.94	0.97	1	0.94	0.97
John Hale	1	0.67	0.8	1	0.65	0.81
Jose M. García	0.97	0.95	0.96	0.98	0.95	0.96
Juan Carlos Lopez	1	0.79	0.88	1	0.79	0.89
Kai Tang	1	0.87	0.93	1	0.88	0.94
Kai Zhang	1	0.85	0.92	1	0.86	0.93

Ke Chen	0.59	0.6	0.59	0.75	0.76	0.76
Keith Edwards	1	0.77	0.87	1	0.75	0.87
Koichi Furukawa	1	0.71	0.83	1	0.72	0.85
Kuo Zhang	1	0.82	0.9	1	0.89	0.94
Lei Chen	0.91	0.77	0.83	0.92	0.82	0.87
Lei Fang	0.74	1	0.85	0.82	1	0.91
Lei Jin	1	1	1	1	1	1
Lei Wang	0.07	0.93	0.13	0.39	0.93	0.6
Li Shen	0.89	0.75	0.81	0.9	0.75	0.82
Lu Liu	0.9	0.91	0.9	0.91	0.94	0.93
M. Rahman	0.91	0.56	0.69	0.94	0.81	0.87
Manuel Silva	0.97	0.94	0.95	0.97	0.95	0.96
Mark Davis	1	0.99	0.99	1	0.96	0.98
Michael Lang	1	0.83	0.91	1	0.9	0.95
Michael Siegel	1	0.82	0.9	0.98	0.78	0.88
Michael Smith	1	0.74	0.85	1	0.85	0.92
Michael Wagner	1	0.38	0.55	1	0.6	0.77
Ning Zhang	0.99	0.73	0.84	0.95	0.81	0.88
Paul Brown	0.51	0.76	0.61	0.73	0.8	0.77
Paul Wang	1	0.92	0.96	1	0.92	0.96
Peter Phillips	1	0.74	0.85	1	0.77	0.88

Philip J. Smith	0.83	0.63	0.72	0.89	0.62	0.74
Ping Zhou	1	0.84	0.91	1	0.9	0.95
Qiang Shen	1	0.88	0.93	1	0.87	0.93
R. Balasubramanian	1	0.47	0.64	1	0.58	0.76
R. Cole	1	0.1	0.17	1	0.35	0.59
R. Ramesh	0.94	0.74	0.83	0.93	0.75	0.84
Rafael Alonso	0.93	0.21	0.34	0.93	0.32	0.54
Rakesh Kumar	1	0.95	0.97	1	0.91	0.95
Richard Taylor	0.99	0.82	0.9	0.97	0.85	0.91
Robert Allen	0.99	0.88	0.93	0.96	0.92	0.94
Robert Schreiber	1	0.36	0.53	1	0.38	0.61
S. Huang	1	1	1	1	1	1
Sanjay Jain	1	0.98	0.99	1	0.96	0.98
Satoshi Kobayashi	0.83	0.36	0.5	0.91	0.44	0.63
Shu Lin	1	0.8	0.89	1	0.8	0.9
Steve King	1	0.31	0.47	1	0.49	0.7
Thomas D. Taylor	1	1	1	1	1	1
Thomas Hermann	0.94	0.8	0.86	0.93	0.85	0.89
Thomas Meyer	1	0.33	0.5	1	0.53	0.73
Thomas Tran	1	0.43	0.61	1	0.5	0.71
Thomas Wolf	0.89	0.32	0.48	0.92	0.51	0.68

Thomas Zimmermann	1	0.85	0.92	1	0.86	0.93
Wei Xu	0.82	0.93	0.87	0.9	0.88	0.89
Wen Gao	0.95	0.97	0.96	0.96	0.97	0.96
William H. Hsu	1	0.66	0.79	1	0.69	0.83
X. Zhang	0.89	0.85	0.87	0.92	0.9	0.91
Xiaoming Wang	1	0.96	0.98	1	0.93	0.97
Xiaoyan Li	1	0.91	0.95	1	0.95	0.97
Yan Tang	0.78	0.94	0.85	0.86	0.91	0.89
Yang Wang	0.29	0.54	0.38	0.69	0.77	0.73
Yang Yu	0.96	0.88	0.92	0.97	0.89	0.93
Yi Deng	0.87	0.97	0.92	0.89	0.98	0.93
Yong Chen	0.78	0.67	0.72	0.83	0.8	0.82
Yoshio Tanaka	1	0.86	0.92	1	0.87	0.93
Young Park	1	0.67	0.81	1	0.81	0.9
Yu Zhang	0.48	0.65	0.55	0.69	0.75	0.72
Yue Zhao	0.93	0.98	0.95	0.95	0.93	0.94
Yun Wang	1	0.49	0.66	1	0.69	0.83
Z. Wang	0.83	0.95	0.88	0.89	0.96	0.92

Table A.27: Percent Agreement Showing the Interrater Reliability.

CSV File#	**Annotator/rater#1**	**Annotator/rater#2**	**Annotator/rater#3**	**%Agreement**
1	1	1	1	100
2	1	1	1	100
3	1	1	1	100
...	...	...	...	...
137	1	1	1	100
Total interrater reliability score				100

Table A.28: Kapa Coefficient Score Interpretation.

Value of Kappa	**Level of Agreement**	**% of Data that are Reliable**
0 – 0.20	No agreement	0 – 4%
0.21 – 0.39	None to slight	4 – 15%
0.40 – 0.59	Fair	15 – 35%
0.60 – 0.79	Moderate	35 – 63%
0.80 – 0.90	Strong	64 – 81%
Above 0.90	Almost perfect agreement	82 – 100%

TABLE A.29: Example Papers Metadata of Author Block "M qadir".

Co-authors	Paper title	Venue	Author affiliation	Author email	Abstract	Keywords	Date
Syed Zubair Ahmad, Mohammad Saeed Akbar, Muhammad Abdul Qadir	High Speed Scalable Mobility Management Architecture over Infrastructural WLAN	IEEE International Multitopic Conference	Center for Distributed and Semantic Computing Mohammad Ali Jinnah University Islamabad Pakistan	aqadirjinnah.edu.pk	Mobility management in a fast moving environment is convoluted...	Fast Mobility Management,Movement Detection and Prediction,Link Layer Triggers,Wireless LAN (WLAN) Hotspots	1/1/2007
Muhammad Fahad, Muhammad Abdul Qadir, Muhammad Wajahat Noshairwan, Nadeem Iftikhar	DKP-OM A Semantic Based Ontology Merger	I-Semantics	Center for Distributed and Semantic Computing Mohammad Ali Jinnah University Islamabad Pakistan	aqadirjinnah.edu.pk	Accurate mapping and merging of multiple ontologies to produce consistent and coherent....	Ontology Mapping and Merging, Disjoint Knowledge Preservation, Ontology Engineering, Knowledge Modelling, Semantic Computing	1/1/2007

Umar Farooq, Antoine Nongaillard, Yacine Ouzrout, Muhammad Abdul Qadir	A Feature-Based Reputation Model for Product Evaluation	International Journal of Information Technology and Decision Making	Department of Computer Science, Capital University of Science and Technology, Islamabad, Pakistan	aqadircust.edu.pk	Knowing the strengths and weaknesses of a product is very important for manufacturers and customers to make decisions...	Product reputation model,product evaluation,reputation system,feature reputation,ratings aggregation	7/1/2016
Umar Farooq, Antoine Nongaillard, Yacine Ouzrout, Muhammad Abdul Qadir	Product reputation evaluation: the impact of conjunction on sentiment analysis	Software, Knowledge, Information Management and Applications	Department of Computer Science, Capital University of Science and Technology, Islamabad, Pakistan	aqadirjinnah.edu.pk	The two main issues in sentiment analysis are word sense disambiguation and conjunction analysis...	sentiment analysis, opinion mining, conjunction analysis, conjunction rules, compound sentences.	1/1/2013

Appendix B

Figures

In Figure B.1 we can see that the paper title as well as the paper venue have more variability among scores and have higher eigenvalues then other features. However, they have been used as supporting evidence in cases where other features needed extra surety to assign the correct author group. Co-authors' names have been used rarely during the manual disambiguation process and have very less variability in their score values.

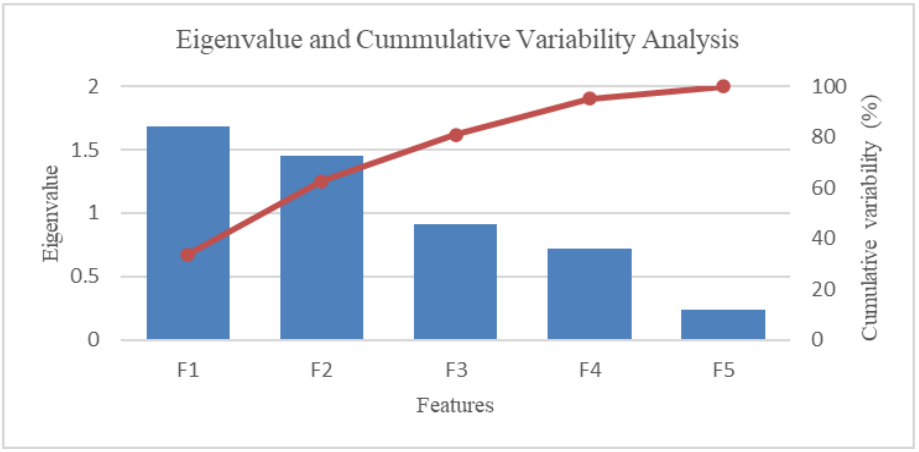

FIGURE B.1: Eigen Values of Features versus their Cumulative Variable Percentage.

Two main reasons for not using co-authors' name more frequently during the manual disambiguation process, is: majority of the cases are resolved using author affiliation and email features, and, manual matching of co-authors' names is tiresome, as humans prefer to analyze fewer challenging features. This can be a biased fact, as SFS-based feature selection has proved that co-author

name is very useful either used alone or in combination with other features. Also, among positively correlated features i.e., "co-authors name, paper title, and paper venue", co-authors' name can be considered a better option due to the overall impact on the pF1 score.

Moreover, features that are given high scores during the manual disambiguation process have given low eigenvalues as well as low individual variability scores, e.g., author affiliation and email. This is because, they have less variability in score values, yet are the most powerful of all the analyzed features for assigning the correct author group.

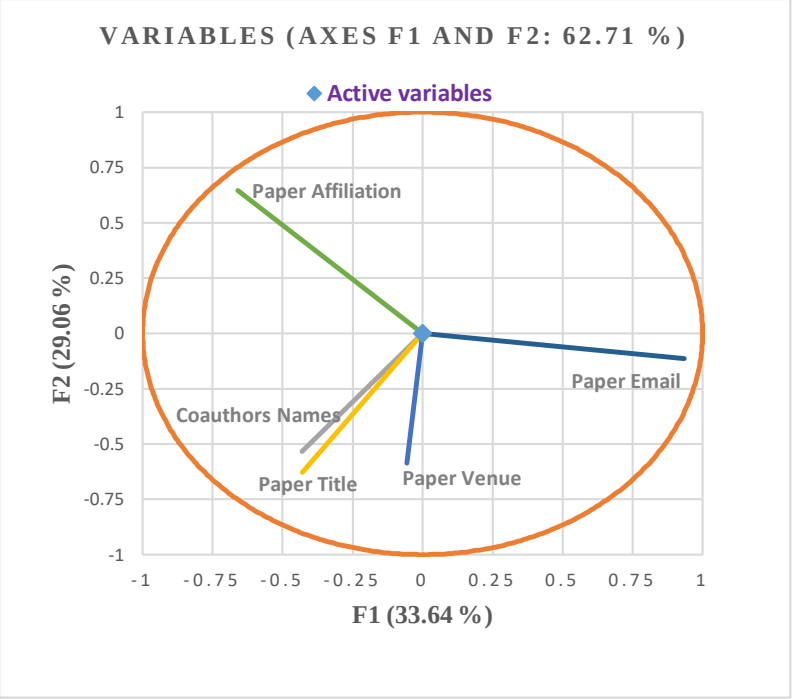

FIGURE B.2: Correlation Circle.

Figure B.2 represents the correlations between variables where the information is interpreted in terms of angles between variables or variables and PCA components (axis). Narrow angles represent positively linked variables like "co-authors name with respect to paper title and venue", right angles represent variables which are unrelated to each other like "authors affiliation with respect to co-authors name and paper title", obtuse angles represent negative relationships like "author affiliation with respect to author email and paper venue", and vector lengths represent the quality in the investigation PCA dimension (which is sufficient in our case).

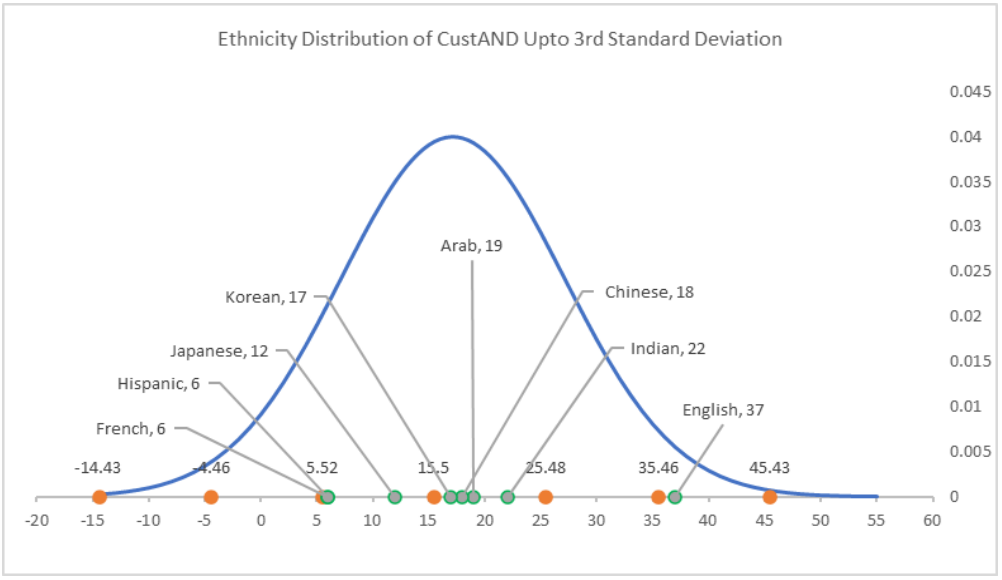

FIGURE B.3: Ethnicity Distribution of CustAND with Respect to Third Standard Deviation.

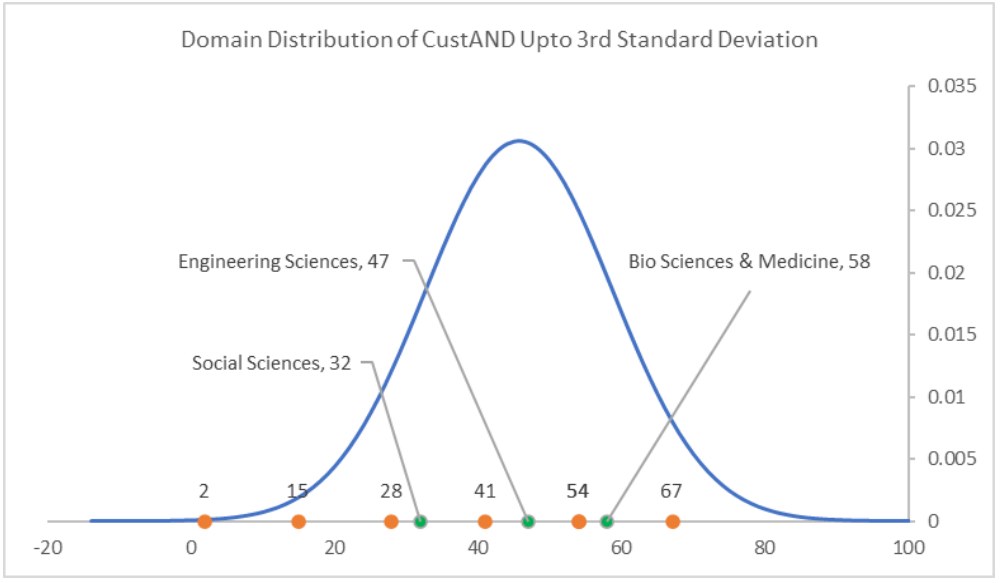

FIGURE B.4: Domain Distribution of CustAND up to Third Standard Deviation.

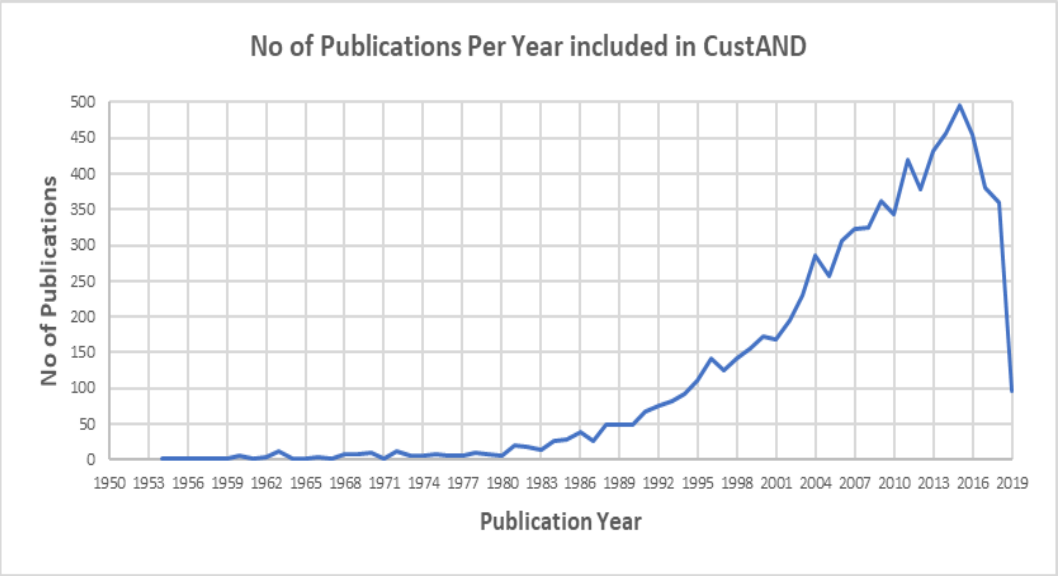

FIGURE B.5: Number of Publications Per Year Included in CustAND Collection.

FIGURE B.6: Same Name Distinct Authors per Ambiguous Block with Respect to their Publications Count.

Appendix C

Equations

C.1 Cohen's Kappa Metric

Cohen's kappa metric is used to assess the agreement between two raters, who each classify items into mutually exclusive categories [33].

$$\kappa = \frac{pr_{(a)} - pr_{(e)}}{1 - pr_{(e)}} \tag{C.1}$$

$$pr_{(e)} = \frac{(\frac{cm^1 * rm^1}{n}) + (\frac{cm^2 * rm^2}{n})}{n} \tag{C.2}$$

Where: cm^1 represents column 1 marginal (row and column intersecting cells), cm^2 represents column 2 marginal, rm^1 represents row 1 marginal, rm^2 represents row 2 marginal, and n represents the number of observations.

Appendix D

Example

D.1 Handling Multi-Author Papers

MHCF addresses multi-author papers by allowing each paper to belong to multiple clusters without any impact on the precision, recall, and F1 scores. Here's how it works:

Consider a set of papers $P = \{p_1, p_2, p_3\}$:

- Paper p_1: Authored by Author A and Author B.
- Paper p_2: Authored by Author A and Author B.
- Paper p_3: Authored by Author B.

D.1.1 For Author A

- Similarity Calculation:
  - $F_{\text{similarity}}(p_1, p_2) \geq$ threshold (both papers are authored by Author A).
  - $F_{\text{similarity}}(p_1, p_3) <$ threshold (no match for Author A).
- Clustering:
  - Papers p_1 and p_2 will be clustered together under Author A's cluster c_A.
  - Paper p_3 will not be included in Author A's cluster c_A.

D.1.2 For Author B

- Similarity Calculation:
  - $F_{\text{similarity}}(p_1, p_2) \geq$ threshold (both papers are authored by Author B).
  - $F_{\text{similarity}}(p_1, p_3) \geq$ threshold (both papers are authored by Author B).
- Clustering:
  - Papers p_1 and p_2 will be clustered together under Author B's cluster c_B.
  - Paper p_3 will also be included in Author B's cluster c_B.

D.1.3 Effect on Precision, Recall, and F1 Scores

The framework ensures that clusters for each author are formed independently:

- For Author A, only papers p_1 and p_2 are considered, and metrics are calculated based on this cluster.
- For Author B, papers p_1, p_2, and p_3 are considered, and metrics are calculated based on this cluster.

Since the clustering process and subsequent evaluation for each author are handled separately, the presence of multi-author papers does not affect the precision, recall, and F1 scores for each author's cluster.