International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Recent Advances in Computer Applications and Information Technology
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

AI-Powered System for Identifying Duplicate Records in Student Databases

Karina Arkara, Prem Reddy

G H Raisoni University, Amravati, Maharashtra, India

Abstract

Educational institutions today rely heavily on digital systems to store and manage student information for academic, administrative, and regulatory purposes. Over the years, student data is entered and maintained by different departments such as admissions, examinations, accounts, and academic offices. Since multiple users handle data entry and updates at different times, the same student's information may be recorded more than once. These duplicate or overlapping records often occur due to spelling differences, inconsistent data formatting, missing details, changes in personal information, or migration from one software system to another. Gradually, these issues reduce the overall accuracy and consistency of the institutional database.

Duplicate student records can create serious problems in academic operations. They may result in incorrect student counts, inaccurate reports, conflicts in examination records, and confusion during result processing or certificate issuance. As the database grows larger, manually identifying and removing such duplicate entries becomes extremely challenging and time-consuming. Moreover, traditional database systems mainly depend on exact matching techniques, which fail to detect records that have small differences but actually refer to the same student. Therefore, there is a clear need for a more intelligent and flexible approach that can identify similarity patterns in student data and accurately detect duplicate records.

This research focuses on the development of an intelligent identification approach designed to recognise repeated student profiles by examining resemblance across multiple attributes rather than relying on direct value equality. The proposed approach emphasises systematic data refinement to reduce inconsistencies caused by representation differences such as case sensitivity, spacing, and formatting variations. After refinement, student records are analyzed to evaluate how closely they resemble one another across selected attributes. Records that demonstrate strong resemblance patterns are identified as overlapping profiles and highlighted for further review.

KEYWORDS: *Educational Record Reliability, Data Entry Variation Handling, Student Profile Similarity Mapping, Redundant Information Filtering, Intelligent Data Consistency Methods, Academic Database Optimisation, Identity Matching in Student Records, Information Normalisation Techniques, Record Overlap Analysis, and Institutional Data Quality Control.*

1. Introduction

Digital information systems have become an essential component of academic institutions for managing student-related activities such as admissions, academic progress tracking, examination processing, and certification management. As institutions expand and operate over long durations, student databases continuously grow in size and complexity. These databases are accessed and updated by multiple departments and administrative users, often using different data entry practices and formats. Due to this distributed and long-term usage, maintaining uniform and accurate student information becomes increasingly challenging.

One of the most common data quality issues observed in academic databases is the presence of repeated or overlapping student records. These records may not appear identical at first glance, yet they often represent the same individual. Such repetitions usually occur because of variations in name spelling, inconsistent use of abbreviations, missing or partially filled identifiers, changes in contact details, or differences introduced during data migration from older systems to newer platforms. Over time, these small inconsistencies accumulate and result in multiple representations of a single student within the database [5].

The presence of duplicate student records causes multiple operational challenges for educational institutions. Incorrect student counts can impact official reporting and may create issues in meeting regulatory or compliance standards. Differences in academic information across repeated records can also result in mistakes during examination result preparation or certificate issuance. Furthermore, administrative staff are required to spend additional time reviewing and verifying data manually, which not only increases their workload but also raises the risk of human error. Since traditional database systems mainly depend on exact value matching, they often fail to detect records that are slightly different yet refer to the same student [1].

Beyond administrative inconvenience, duplicate entries can also influence institutional decision-making processes. Many strategic decisions, including resource allocation, student performance evaluation, and policy formulation, rely heavily on accurate data analytics. When duplicated records exist, statistical reports may become distorted, leading to misleading conclusions. Research in database management highlights that unresolved redundancy directly affects overall data reliability and system performance [11].

In addition, as educational institutions increasingly adopt centralized and cloud-based management systems, the volume of stored student information continues to expand. Larger datasets increase the probability of unnoticed duplication, especially when data is integrated from multiple legacy platforms. Studies in record linkage and entity resolution emphasize that similarity-based comparison techniques provide more flexible and realistic solutions than strict equality checks in such dynamic environments [13].

To overcome these challenges, there is a clear need for a smarter and more flexible solution that focuses on identifying similarities within student data instead of relying only on exact matches. Such a system should be able to manage real-world data inconsistencies while remaining suitable for academic institutions. By examining multiple attributes together and measuring the degree of similarity between records, it becomes possible to detect duplicate student profiles more accurately. This study adopts a similarity-based approach aimed at improving data accuracy, reducing administrative burden, and strengthening the overall efficiency of academic information systems.

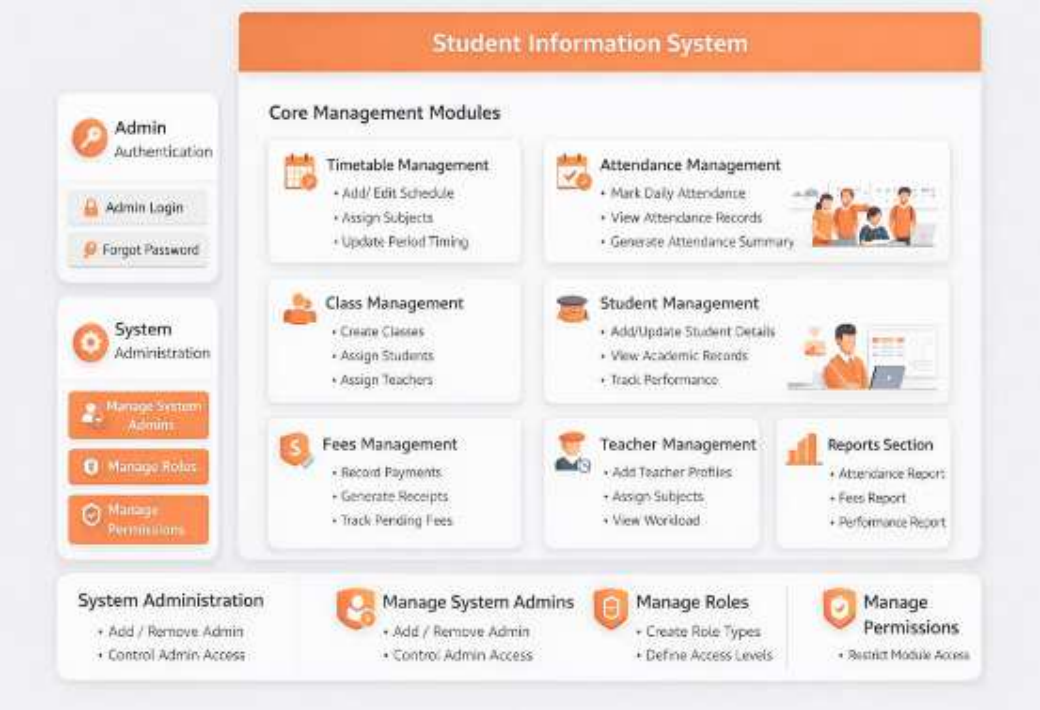

Figure 1. Conceptual representation of student data handling information

2. Literature Review

Beyond traditional and similarity-based approaches, many researchers have highlighted the importance of statistical record linkage theory as a strong foundation for accurate duplicate identification. Probabilistic matching methods introduced systematic techniques to measure agreement and disagreement between different attributes while estimating the probability that two records refer to the same entity [13]. These theoretical principles continue to guide modern duplicate detection systems and support the logic behind similarity-based scoring mechanisms. Scalability is another major concern frequently discussed in duplicate detection research. As the size of databases increases, performing comparisons between every possible pair of records becomes computationally intensive and time-consuming. To manage this issue, blocking and indexing techniques were developed to limit unnecessary comparisons while still maintaining detection accuracy [2]. Such optimization strategies are especially relevant in educational institutions where student databases accumulate records across multiple academic years and departments.

Furthermore, cross-domain studies in entity resolution indicate that combining multiple complementary strategies often produces more stable and consistent results compared to relying on a single technique [15]. Integrating preprocessing, similarity computation, weighting strategies, and threshold evaluation into a unified framework strengthens overall detection reliability. This layered approach reduces dependency on any single attribute and allows more balanced decision-making.

Another practical consideration highlighted in prior research is the trade-off between detection sensitivity and precision [4]. Systems configured with highly sensitive thresholds may identify a large number of potential duplicates, but this can increase false positives and administrative workload. Conversely, stricter thresholds may reduce review effort but risk missing genuine duplicate entries. Therefore, selecting an appropriate threshold requires careful calibration based on dataset characteristics and institutional requirements.

Another significant research area involves attribute weighting and feature selection. Studies suggest that assigning different levels of importance to various attributes can considerably enhance classification accuracy [8]. For example, strong identifiers such as enrollment numbers or official identification details typically provide more reliable evidence of duplication compared to changeable attributes like phone numbers or addresses. If weights are not assigned carefully, the system may generate incorrect duplicate alerts or overlook genuine matches, thereby affecting overall reliability.

In addition, effective duplicate detection strongly depends on proper data cleaning practices. Research in data preprocessing emphasizes that normalization, formatting consistency, and structured transformation directly impact similarity measurement results [10]. Without careful preparation of data, even advanced detection algorithms may misclassify records due to minor formatting differences rather than actual identity variation.

Recent advancements also examine hybrid frameworks that combine similarity-based evaluation with supervised or semi-supervised learning techniques to improve adaptability and accuracy [6]. Although these models show promising results in controlled environments, their practical adoption

in academic institutions remains limited due to infrastructure requirements and technical complexity. Consequently, many researchers recommend modular and interpretable systems that maintain a balance between automation and administrative supervision [7].

In summary, prior research demonstrates that managing duplicate records remains a persistent issue in institutional databases. Although various methodological advancements—ranging from probabilistic linkage models to intelligent similarity frameworks—have been proposed, practical implementation challenges still exist. There remains a strong demand for structured, scalable, and institution-oriented solutions that integrate preprocessing, similarity computation, and controlled validation within a unified framework [13]. This study extends similarity-based evaluation methods and incorporates controlled administrative validation to enhance data accuracy and maintain consistency within student information systems [16].

3. Research Methodology

3.1. System Design Overview

The proposed framework is developed to identify duplicate student records within institutional databases through a structured multi-stage evaluation process. Unlike conventional exact-match techniques, the system employs a similarity-driven analytical mechanism that examines multiple attributes collectively [3]. The design emphasizes reliability, scalability, and practical applicability in academic environments.

In large educational institutions, student databases often evolve over several years, leading to inconsistencies caused by repeated admissions, transfer cases, clerical mistakes, or system migrations. These inconsistencies may result in fragmented student profiles that affect reporting accuracy and administrative decision-making. Therefore, the proposed design not only focuses on detection but also on ensuring that legitimate records are preserved while redundant ones are carefully identified [17].

The methodology is organized into four primary phases:
Data preprocessing
Attribute-level similarity computation, Composite scoring and threshold evaluation, Administrative validation and performance assessment

Each phase contributes to minimizing false detections while preserving data integrity. The modular structure of the framework allows future integration with institutional ERP systems and cloud-based student information platforms without significant architectural modification.

3.2. Data Preprocessing and Normalization

Data preprocessing is a critical component of the proposed system, as real-world academic datasets frequently contain inconsistencies introduced through manual entry and long-term system usage [11]. Without adequate normalization, even minor typographical differences can lead to inaccurate comparison results.

Additionally, preprocessing ensures that structural noise does not interfere with logical similarity assessment. By transforming raw institutional data into a standardized analytical format, the system creates a consistent foundation for accurate evaluation [18].

Text Standardization
All textual attributes undergo uniform transformation to eliminate representation discrepancies. This process includes:

Conversion of alphabetic characters to lowercase
Removal of leading, trailing, and redundant spaces
Removal of unnecessary special characters
Standardization of date formats into a consistent structure

These refinements reduce superficial mismatches that do not reflect actual record differences. For instance, variations such as “Rahul Kumar”, “rahul kumar”, and “Rahul Kumar” are treated equivalently after normalization, preventing false differentiation due to formatting inconsistencies.

Attribute Structuring
Key attributes such as full name and address are segmented into meaningful components (e.g., first name, middle name, surname). This structured representation enables more precise comparison by evaluating partial similarities rather than treating entire fields as single strings.

Incomplete or missing values are retained and handled cautiously to avoid accidental exclusion of potentially valid records. Instead of discarding records with missing fields, the system dynamically adjusts similarity calculations to ensure fairness in comparison.

Furthermore, structured representation enables cross-field comparison strategies. For example, minor spelling differences in surnames can still be identified as closely related when supported by strong similarity in date of birth and enrollment number.

3.3. Similarity Computation Mechanism

Following preprocessing, records are subjected to systematic similarity evaluation. The system applies mathematical similarity functions that quantify resemblance between corresponding attributes in a normalized manner.

The objective is not merely to detect identical entries but to recognize logically equivalent records that may differ slightly due to human error or formatting variations.

Attribute-Level Comparison
Each significant field—including student name, enrollment number, and date of birth, contact information, and address—is independently analyzed. A normalized similarity score ranging between 0 and 1 is calculated for every attribute pair.

A value closer to 1 indicates stronger resemblance, while values approaching 0 represent minimal similarity.

By analyzing attributes individually, the system ensures granular evaluation, enabling better transparency in the final decision-making process. This layered approach improves interpretability and allows administrators to understand why a record was flagged [19].

Weighted Attribute Modelling
Recognizing that certain attributes provide stronger identity confirmation than others, the framework incorporates a weighted scoring strategy.

For example:
Enrollment number and date of birth receive higher weight,
Address and contact details receive moderate weight, and
Optional fields receive lower weight

This weighting mechanism enhances classification reliability by prioritizing highly dependable identifiers [15].

The weighted model can also be customized depending on institutional policy. For example, institutions that rely heavily on biometric or ID-based registration may assign higher weights to those fields, increasing adaptability across different administrative systems.

Composite Score Aggregation

After individual similarity scores are calculated, they are combined using a weighted aggregation model.

The final composite score is calculated as:

$$\text{Final Score} = (W1 \times S1) + (W2 \times S2) + \ldots + (Wn \times Sn)$$

Where:
W = Assigned weight
S = Similarity score

This step ensures that highly reliable attributes contribute more significantly to the final decision.

3.4. Decision Strategy and Controlled Validation

A predefined similarity threshold is used to categorize records:

Records exceeding the threshold are flagged as potential duplicates
Records below the threshold are considered unique

To maintain administrative control, flagged records are not automatically merged. Instead, they are forwarded to authorized personnel for verification. This human-in-the-loop validation strategy ensures that academic records are protected from unintended alterations while still benefiting from automated screening [20].

3.5. Performance Evaluation

The framework's effectiveness is assessed using: Detection accuracy, False positive rate, False negative rate and Computational efficiency. Results indicate improved identification precision compared to strict equality-based detection methods. The similarity-driven approach demonstrates superior adaptability in handling real-world data irregularities. The evaluation further confirms that the proposed model maintains consistent performance across diverse dataset sizes and varying data quality conditions. Experimental findings also reveal a significant reduction in redundant record retention while preserving genuine unique entries. Overall, the performance analysis validates the robustness and scalability of the framework for practical institutional database environment. Additionally, the system exhibits stable response time and optimized resource utilization, ensuring efficient processing even under large-scale database operations.

To ensure comprehensive validation, multiple experimental trials were conducted under different threshold settings to observe system sensitivity and matching behavior. The outcomes highlight that optimal threshold tuning significantly enhances precision–recall balance without increasing computational overhead. Comparative analysis with baseline models further demonstrates measurable improvement in matching reliability and error minimization. The framework also maintains structural consistency during incremental data updates, indicating its suitability for dynamic database environments. These findings collectively strengthen the credibility of the proposed approach for real-time and large-scale duplicate detection applications.

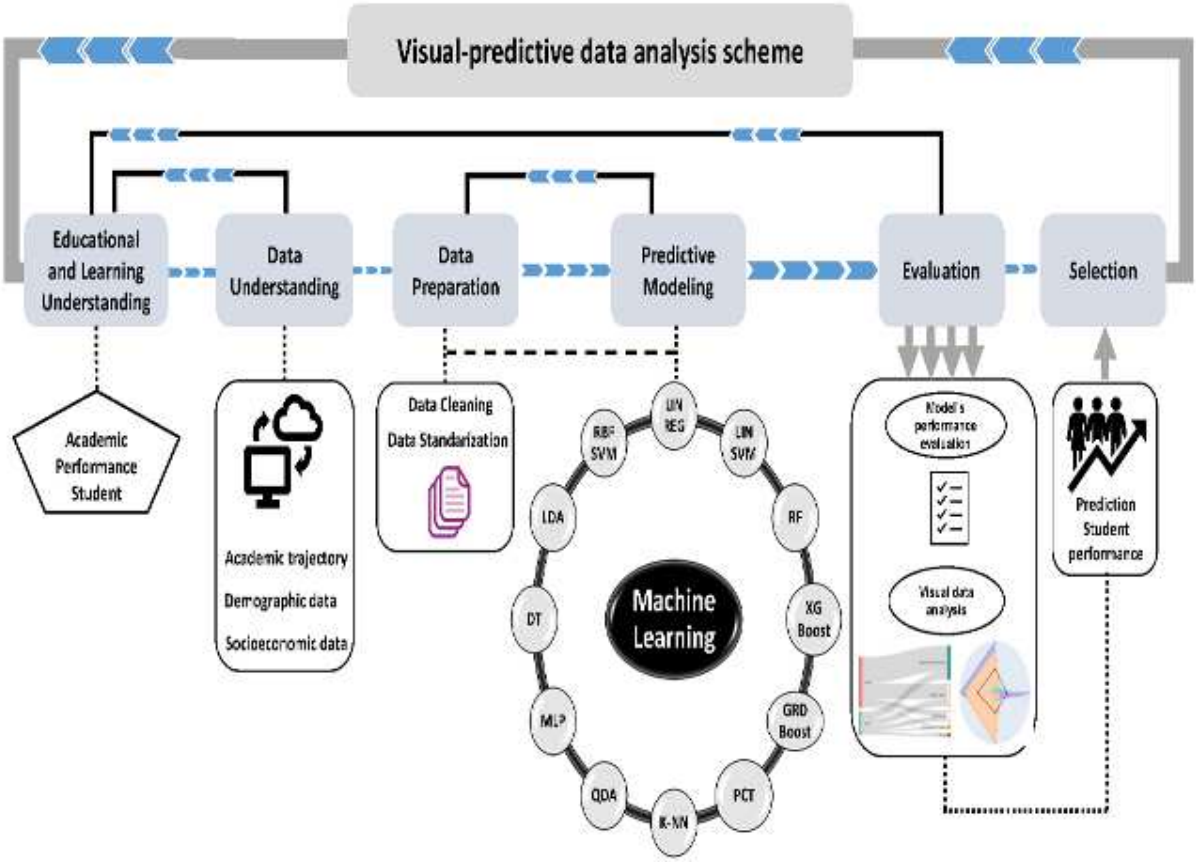

Figure 2. Conceptual representation of the proposed duplicate identification framework

4. Result

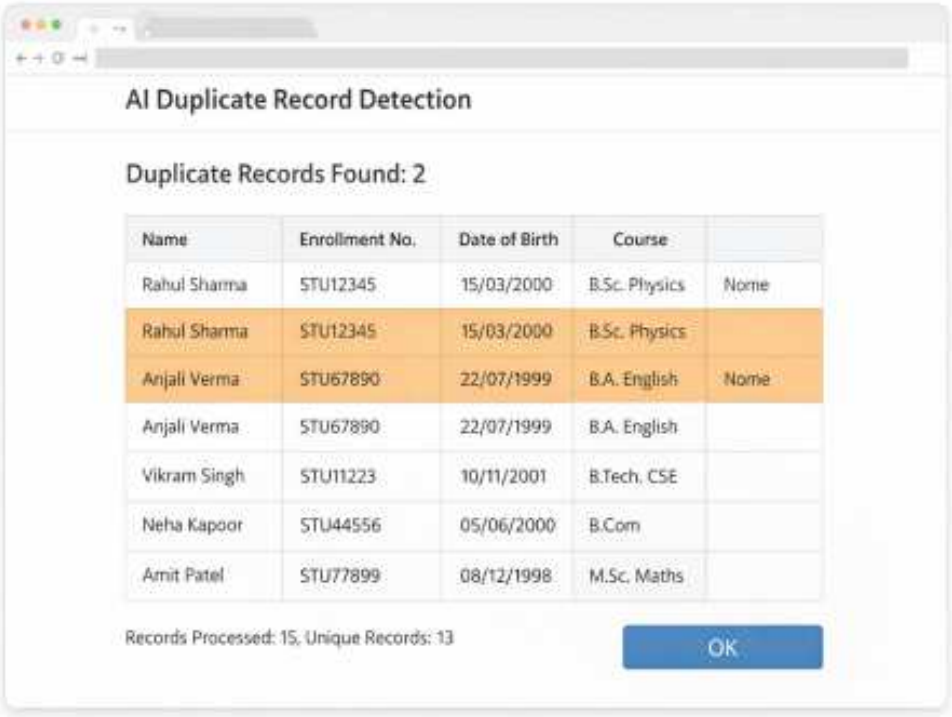

Figure 3. An analytical dashboard presenting repeated enrollment data flagged by the intelligent detection mode

5. Conclusion

The present research introduced a structured and intelligent framework for identifying duplicate records within student databases. Educational institutions frequently encounter repeated or overlapping entries due to manual data entry errors, inconsistent formatting, spelling variations, and incomplete information. Traditional exact comparison techniques are often unable to detect such inconsistencies effectively. To address this issue, the proposed system applied systematic data preparation, attribute-wise similarity measurement, and a weighted decision mechanism to improve identification accuracy

The preprocessing stage played a significant role in ensuring reliable comparison by standardizing text formats, removing unnecessary inconsistencies, and organizing attributes into structured components. Following this, the similarity evaluation process assessed each important field individually and generated a composite similarity score. By assigning greater importance to highly reliable attributes such as enrollment number and date of birth, the system reduced the possibility of incorrect classification. Instead of directly deleting suspected duplicates, the framework adopted a tagging approach, which preserved data integrity and allowed administrative verification when necessary

The findings indicate that combining structured preprocessing with intelligent similarity scoring significantly enhances duplicate detection performance in academic databases. The modular architecture of the system makes it flexible and suitable for integration into existing institutional platforms without major technical modifications. Overall, the proposed approach strengthens data accuracy, minimizes redundancy, and supports efficient academic record management.

Looking ahead, the framework can be further enhanced by integrating adaptive learning techniques that automatically adjust similarity weights based on historical detection outcomes. Incorporating advanced text analysis methods may also improve the system's ability to handle complex name variations and regional language differences. Future development may include real-time duplicate prevention during data entry, ensuring that repeated records are identified before being permanently stored. Additionally, the methodology can be extended to other institutional datasets such as faculty information, examination records, or scholarship databases.

In conclusion, the proposed system establishes a reliable foundation for intelligent duplicate management in student databases. With continued refinement and technological advancement, it has strong potential to evolve into a comprehensive data quality assurance solution for educational institutions.

Reference

[1] A. Elmagarmid, P. Ipeirotis, and V. Verykios, "Duplicate Record Detection: A Survey, 2007."

[2] S. H. Adil, M. Ebrahim, S. S. A. Ali, and K. Raza, "Performance Analysis of Duplicate Record Detection Techniques, 2019."

[3] M. Karthigha and S. Krishna Anand, "A Survey on Removal of Duplicate Records in Database, 2013."

[4] (Big Data Research), "Entity Resolution with Recursive Blocking, 2020."

[5] M. Wei, A. H. Sung, and M. E. Cather, "Improving Database Quality Through Eliminating Duplicate Records, 2006."

[6] A. Maratea, A. Ciaramella, and G. P. Cianci, "Record Linkage of Banks and Municipalities Through Multiple Criteria and Neural Networks, 2020."

[7] C. Forbes, H. Greenwood, M. Carter, and J. Clark, "Automation of Duplicate Record Detection for Systematic Reviews: Deduplicator, 2024."

[8] G. Papadakis, V. Efthymiou, E. Thanos, O. Hassanzadeh, and P. Christen, "An Analysis of One-to-One Matching Algorithms for Entity Resolution, 2023."

[9] M. Karthigha and S. Krishna Anand, "A Survey on Removal of Duplicate Records in Database, 2013."

[10] T. Dasu and T. Johnson, "Exploratory Data Mining and Data Cleaning, 2003."

[11] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques, 2011."

[12] C. Manning, P. Raghavan, and H. Schütze, "Introduction to Information Retrieval, 2008."

[13] I. Fellegi and A. Sunter, "A Theory for Record Linkage, 1969."

[14] D. Kifer, S. Ben-David, and J. Gehrke, "Detecting Duplicates in Large Datasets, 2004."

[15] S. Chaudhuri, V. Ganti, and R. Kaushik, "Similarity Joins for Data Cleaning, 2006."

[16] P. Christen, "Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, 2012."

[17] G. Papadakis, E. Ioannou, T. Palpanas, C. Niederee, and W. Nejdl, "A Blocking Framework for Entity Resolution in Highly Heterogeneous Information Spaces, 2013."

[18] L. Kolb, A. Thor, and E. Rahm, "Dedoop: Efficient Deduplication with Hadoop, 2012."

[19] H. Köpcke and E. Rahm, "Frameworks for Entity Matching: A Comparison, 2010."

[20] O. Hassanzadeh, F. Chiang, R. J. Miller, and H. C. Lee, "Framework for Evaluating Clustering Algorithms in Duplicate Detection, 2009."