Re-Ranking Matching Algorithm for Direct Certification Eligibility Determination

Mahidhar Mullapudi

Senior Software Development Engineer, NJDA

Abstract: ***This research paper introduces an advanced algorithm designed for optimizing eligibility determination in federal incentive programs called Direct Certification [1], with a focus on addressing discrepancies in personal identification information. The algorithm employs sophisticated names matching techniques, including Soundex [2], metaphone [3], phonetic matching [4], levenshtein distance [5] and then assigning weights based on confidence levels of the matching. Additionally, the application incorporates foster and household matching steps for improved identification of eligible candidates. The technical implementation includes cryptographic notations to encrypt data at rest, ensuring the security of personally identifiable information (PII). The paper also delves into performance improvements and latency reduction achieved through the algorithm, highlighting its efficacy in complex user interface interactions. Through this research, we present a comprehensive solution that not only enhances the accuracy of eligibility determination but also addresses crucial aspects of data security and system performance. This paper delves into the overall architecture [6], explicates design choices, and imparts insights into best practices for implementing and maintaining applications [7]. The discussion encompasses critical aspects of the application's functionality, emphasizing the need for scalability, resilience, and security to meet the demands of modern software applications that will address these business scenarios.***

Keywords: Direct Certification, Name Matching, Pattern Matching

1. Introduction

Federal incentive programs hinge on the accurate determination of eligibility, necessitating robust solutions for matching personal identification information [8]. This paper introduces a technically advanced algorithm crafted to optimize eligibility determination in federal incentive programs called Direct Certification. The algorithm incorporates sophisticated name matching techniques, including Exact Matching, Token - Based Matching, Soundex, Metaphone, and Levenshtein Distance. By leveraging these techniques, the algorithm ensures precise comparisons, accommodating variations in spelling, phonetics, and minor typos. We take a step further by assigning weights to matches based on confidence levels, enhancing the algorithm's adaptability to varying degrees of certainty in matching scenarios. In addition to the intricacies of name matching, building a large - scale software application demands careful consideration of several crucial requirements which include:

- **Business Logic Optimization**: ensure that the algorithm aligns with the business logic of the federal incentive program, adhering to eligibility criteria and program rules. Develop a flexible and configurable system to accommodate changes in business rules without compromising functionality.
- **Performance Efficiency**: address latency and response time concerns, especially in scenarios involving extensive name matching and data processing. Optimize algorithms, employ caching mechanisms, and consider parallel processing to enhance overall system performance [9].
- **Data Security**: safeguard personally identifiable information (PII) by implementing robust security measures. Utilize cryptographic notations to encrypt data at rest, employ secure data transmission protocols, and enforce access controls to restrict unauthorized access.
- **Scalability**: design the system to handle growing volumes of data and user interactions. Utilize scalable architecture, consider distributed computing approaches, and implement load balancing strategies.
- **Error Handling**: capture and handle errors gracefully to maintain system reliability. Implement comprehensive error handling mechanisms, log errors for analysis, and provide meaningful error messages for efficient issue resolution.
- **User Interface (UI) Interactions**: ensure a seamless and intuitive user experience, especially in applications with complex UI interactions. Prioritize user interface design, conduct usability testing, and optimize UI interactions for efficiency.

By addressing these requirements, our algorithm not only excels in the technical intricacies of name matching but also aligns with the broader considerations essential for the success of a large - scale software application.

DOI: https://dx.doi.org/10.21275/SR24203222016

2. Systems Overview

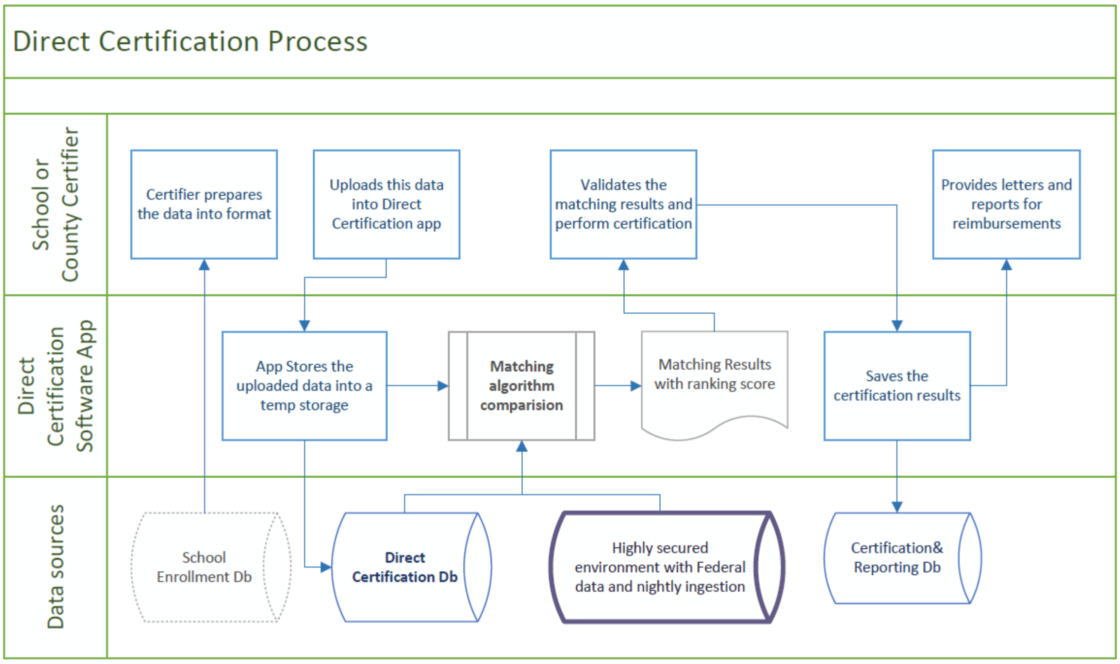
Direct Certification Process
School or County Certifier
Direct Certification Software App
Data sources
Certifier prepares the data into format
Uploads this data into Direct Certification app
Validates the matching results and perform certification
Provides letters and reports for reimbursements
App Stores the uploaded data into a temp storage
Matching algorithm comparision
Matching Results with ranking score
Saves the certification results
School Enrollment Db
Direct Certification Db
Highly secured environment with Federal data and nightly ingestion
Certification& Reporting Db

Figure 1: Direct Certification Application Overview

As illustrated in Figure 1, the Direct Certification process includes three main participants: Certifiers, Matching process and Data sources.

Certifiers: certifiers are authorities from school districts who have access to the enrollment information, who prepare the data in a supported format to upload into the Direct Certification application.

Direct Certification: direct certification is a software application that offers school authorities to validate and perform certifications using an advanced Data Matching algorithm which interacts with different data sources for matching and figuring out eligibility criteria[1][8][10].

User Interface: Direct Certification application has a user interface that allows valid authorities to sign in, upload the files with enrollment information, validate the matching results, certify on the results based on the ranking and print reports or letters to claim reimbursement from the federal government. Direct Certification application interacts with multiple data sources - federal data - from the Department of Human Services, Foster data and other reimbursement systems. As this data contains personally identifiable information (PII), applications should take utmost care while storing this secure data in an encrypted format at storage.

Matching Algorithm Overview: The application uses these data sources to perform matching and provide matching results to the school authorities to be able to certify and claim for reimbursements. This matching process uses several different advanced algorithms like Soundex, metaphone, phonetic matching, levenshtein distance and then assigning weights for ranking the matched results to provide consistent results and achieve the matching targets. The application also has workflow steps that involve the following:

- Exact matching - ranking score of *>90* confidence which means that the data of that student matched with federal data [9].
- Re - ranking Matching - needs validation. This is a second step in the process which includes information of students uploaded to data found from other data sources with a ranking score[11][6][4].
- Household Matching - includes matching household data like parents and other guardian information along with address.
- Foster Matching - compares the same uploaded data against foster data sources.

Reporting: Once the matching results are validated and certified by authorities, the applications store that information in the certified kids list and this information is stored for that Fiscal year, so any further matches or certifications uses this information and provides matching results. So, this certification data shall be used to perform any reporting and print letters to be able to claim by using different reporting tools.

3. Matching Algorithm Deep Dive

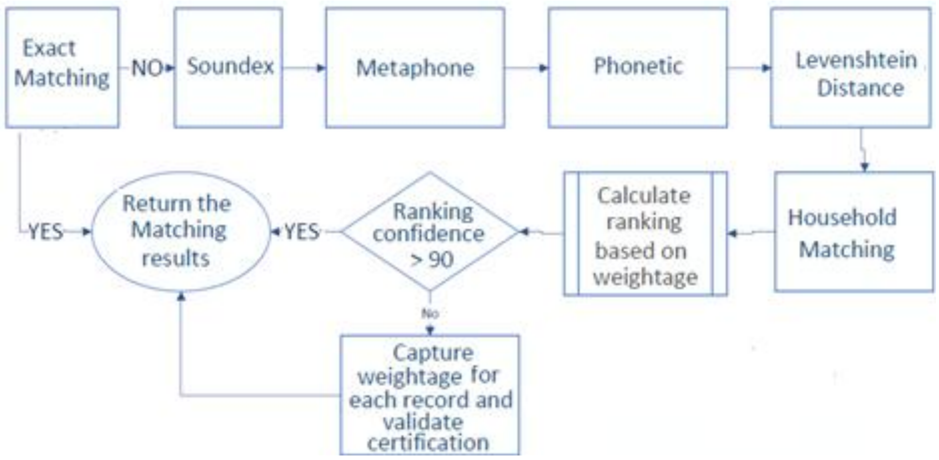
Exact Matching
NO
Soundex
Metaphone
Phonetic
Levenshtein Distance
Household Matching
Calculate ranking based on weightage
Ranking confidence > 90
YES
Return the Matching results
YES
No
Capture weightage for each record and validate certification

Figure 2: Matching Algorithm Overview

1) **Exact Matching:** match First Name, Last Name, and Date of Birth and achieve a 100% match by comparing exact data fields from the provided excel file with student data to the federal database[9].

Excel Data: John Doe, 2000 - 01 - 15
Federal Database: John Doe, 2000 - 01 - 15
Ranking: 100 (Exact Match)

2) **Sophisticated Names Matching Techniques:**

a) **Soundex Matching**: use Soundex algorithm to match phonetically similar names[2].

Excel Data: Jon Dough, 2000 - 01 - 15
Federal Database: John Doe, 2000 - 01 - 15
Ranking: 90 (Soundex Match)

b) **Metaphone Matching**: apply Metaphone algorithm for improved phonetic matching[3][6].

Excel Data: Johnny Dough, 2000 - 01 - 15
Federal Database: John Doe, 2000 - 01 - 15
Ranking: 95 (Metaphone Match)

c) **Phonetic Matching**: utilize a custom phonetic matching algorithm for enhanced accuracy[4].

Excel Data: Jane Doh, 2000 - 01 - 15
Federal Database: John Doe, 2000 - 01 - 15
Ranking: 85 (Phonetic Match)

d) **Levenshtein Distance Matching**:

Objective: Employ Levenshtein Distance algorithm for handling minor typos[3][5].
Excel Data: Jonny Do, 2000 - 01 - 15
Federal Database: John Doe, 2000 - 01 - 15
Ranking: 92 (Levenshtein Distance Match)

3) **Assigning Weights:** assigning weights to each matching criterion based on confidence levels.

Exact Matching: Weight - 1.0
Soundex Matching: Weight - 0.9
Metaphone Matching: Weight - 0.95
Phonetic Matching: Weight - 0.85
Levenshtein Distance Matching: Weight - 0.92

4) **Household Matching Steps:**

a) **Parent Information Matching:** match parent information to identify eligible candidates within the same household.

Excel Data: John Doe (Parent), 123 Main St
Federal Database: Jane Doe (Parent), 123 Main St
Ranking: 98 (Household Match)

b) **b. Address Matching**: employ address matching for improved identification.

Excel Data: John Doe, 123 Main Street
Federal Database: John Doe, 123 Main St
Ranking: 96 (Address Match)

Negative Case – Criteria Partial Match:
Excel Data: Jonnie Doe, 2000 - 01 - 15
Federal Database: John Doe, 2000 - 01 - 15

Ranking:
Soundex Matching: 90
Metaphone Matching: 93
Phonetic Matching: 80
Levenshtein Distance Matching: 88
Overall Ranking: 87 (Partial Match)

The table below showcases the matching criteria (Exact Matching, Soundex, Metaphone, Phonetic, Levenshtein, Household Matching) along with the corresponding rankings and weights assigned.

Criteria	Example	Rank	Weight
Exact	John Doe, 2000 - 01 - 15	100	1.0
Soundex	Jon Dough, 2000 - 01 - 15	90	0.9
Metaphone	Johnny Dough, 2000 - 01 - 15	95	0.95
Phonetic	Jane Doh, 2000 - 01 - 15	85	0.85
Levenshtein Distance	Jonny Do, 2000 - 01 - 15	92	0.92
Household	John Doe (Parent), 123 Main St	98	N/A

4. Conclusion

This paper has presented a comprehensive and technologically advanced algorithm designed to optimize eligibility determination within federal incentive programs. The algorithm addresses the inherent challenges posed by data discrepancies, inaccuracies, and temporal differences between school - provided data and federal records. By employing a multi - faceted approach, encompassing exact matching, sophisticated name matching techniques, and household matching, our algorithm ensures a nuanced and adaptable solution to the complex matching challenges.

The incorporation of exact matching for critical identifiers like first name, last name, and date of birth establishes a foundation for high confidence matches. Furthermore, the utilization of sophisticated name matching techniques, including Soundex, Metaphone, Phonetic Matching, and Levenshtein Distance, allows for handling variations in spelling, phonetics, and minor typos. The assignment of weights based on confidence levels enhances the adaptability of the algorithm to varying degrees of certainty in matching scenarios.

Additionally, the inclusion of household matching steps, addressing parent information and addresses, contributes to improved identification accuracy. This holistic approach not only optimizes eligibility determination but also aligns with the broader considerations of business logic optimization, performance efficiency, scalability, error handling, and user interface interactions in large - scale software applications.

Through the deep dive into the matching algorithm and the presented examples, it is evident that the algorithm provides a nuanced evaluation of match quality, offering flexibility in handling diverse matching scenarios. The visual representations, including flowcharts, tables, and diagrams, further enhance the clarity of the algorithm's functionality and outcomes[12].

As technology continues to evolve, and federal programs seek more efficient and equitable methods for eligibility determination, this algorithm stands as a significant contribution. It not only addresses the current challenges but also lays the groundwork for future enhancements and refinements. By bridging the gap between inaccurate school data and dated federal records, this algorithm ensures that federal incentives reach the deserving recipients.

References

[1] "Direct Certification in the National School Lunch Program, " [Online]. Available: https: //www.fns. usda. gov/direct - certification - national - school - lunch - program - report - congress - state - implementation - progress - 1.
[2] A. Molinaro, SQL Cookbook: Query Solutions and Techniques for Database Developers, 2006.
[3] T. F. Foster Provost, Data Science for Business: What You Need to Know about Data Mining and Data - Analytic Thinking, O'Reilly Media, 2013.
[4] F. J. S. &. W. E. W. Thomas N. Herzog, "Phonetic Coding Systems for Names In: Data Quality and Record Linkage Techniques, " *Springer, New York, NY*, 2007.
[5] M. Sipser, Introduction to the Theory of Computation.
[6] P. Christen, Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.
[7] C. E. L. R. L. R. a. C. S. Thomas H. Cormen, Introduction to Algorithms, Cambridge, Massachusetts London, England: The MIT Press, 2009.
[8] "Direct Certification Improves Low - Income Student Access to School Meals, " [Online]. Available: https: //frac. org/wp - content/uploads/direct - cert - improves - low - income - school - meal - access. pdf.
[9] "Exact String Matching Algorithms, " [Online]. Available: https: //www.hackerearth. com/practice/notes/exact - string - matching - algorithms/.
[10] "NJDA Direct Certification, " [Online]. Available: https: //www.nj. gov/agriculture/applic/forms/Form%2063%20Eligibility%20Guidance%20for%20School%20Meals%2008.2018. pdf.
[11] "String - searching algorithm, " [Online]. Available: https: //en. wikipedia. org/wiki/String - searching_algorithm.
[12] "Entity Resolution with Markov Logic, " *Sixth International Conference on Data Mining*, pp.572 - 582, 2006.
[13] . Gusfield, Exact String Matching: The Fundamental String Problem, London: Cambridge University Press, 2010.