Complete Guide to Fuzzy/Probabilistic Data Matching and Entity Resolution

Introduction

Fuzzy or probabilistic data matching and entity resolution are fundamental processes in data management and analytics. They involve identifying and linking records that refer to the same entity but may have variations due to errors, abbreviations, or inconsistencies. This comprehensive guide delves into the various aspects of fuzzy matching and entity resolution, including different data domains, business use cases, algorithms, and the significance of no-code solutions.

[1]The average company uses more than 400 unique datasets daily.

[1]By 2025, data centric workloads are expected to grow over 2.65 times compared to 2018.

In the world of business, we're always pulling and using data from multiple systems. This also means that we often have to unify 'mismatching' data, and that many times we're adding relational information to our business systems without noticing the relationships. This could be different people in the same household or at the same company, it could be the same person or company with different details, or maybe it's product or address information.

If the information were 'exactly the same' your business systems would probably identify this for you and either update the original record or create some relational indexing key. New information would be 'linked' to existing information.

But because the information isn't exactly the same, it's often treated as new information, which means different things depending on the context, but it essentially means you have fragmented, 'mismatching' and/or 'duplicate' information.

For an end user of any system this can mean multiple searches in one or even multiple systems,

trying to find the right information, and trying to see the full picture.

This could be a customer.

This could be an employee.

Or it could be your CEO.

This could be a doctor, a patient, a paramedic, or a law enforcement officer responding to a call.

This is one of the more common reasons we talk about data quality. This is one of the more common reasons we cannot easily compare information from one system to another system, and this is also one of the more common reasons we replace business systems.

More importantly, this is also one of the more common reasons why operational costs will continue to increase for companies that are less 'data literate', why customers churn, why employees get burnt out, and why strategic initiatives fail to deliver their targeted business benefits.

There are millions of different ways that this can happen and we can't expect perfection but we can do a lot better than we have done historically, by simply understanding the issues and being more proactive. Match Data Pro is a world-class data matching solution, designed to simplify this work.

Table of Contents

1. Different Data Domains

Company Data (Prospects, Customers, Vendors)

> **Challenges:** Companies can have variations in names, such as abbreviations, misspellings, or different legal names. This can lead to fragmented data and hinder accurate analysis.
>
> **Solution:** Fuzzy matching algorithms can identify and link these variations to a single entity, enabling a unified view of company data.

Contact Data (Consumer and B2B Contact Data)

> **Challenges:** Contact information like names, phone numbers, and email addresses often have variations due to typos, different formats, or aliases.

Solution: Fuzzy matching helps in identifying duplicate contacts, merging them, and maintaining a clean and accurate contact database.

Address Data

Challenges: Addresses can have variations in terms of abbreviations, misspellings, or different formats, leading to delivery issues and data inconsistencies.

Solution: Fuzzy matching is crucial for address validation, standardization, and deduplication to ensure accurate geolocation and address matching.

Product Data

Challenges: Product names and descriptions can vary due to abbreviations, misspellings, or synonyms, making it challenging to categorize and analyze products.

Solution: Fuzzy matching algorithms assist in product matching, categorization, and standardization, enabling efficient inventory management and analysis.

Parts Data

Challenges: Parts and components can have variations in names or codes across different datasets, leading to inventory discrepancies and procurement challenges.

Solution: Fuzzy matching is essential for inventory management, part identification, and traceability, ensuring accurate and consistent part tracking.

Materials Data

Challenges: Materials and raw materials can have variations in names or codes, complicating procurement processes and material traceability.

Solution: Fuzzy matching aids in material standardization, categorization, and procurement optimization by linking related records and eliminating duplicates.

Assets Data

Challenges: Assets like equipment or machinery can have variations in names or identifiers, making asset tracking and maintenance management difficult.

Solution: Fuzzy matching algorithms assist in asset identification, tracking, and maintenance scheduling by linking related records and ensuring data consistency.

2. Different Business Use Cases

Data Quality

Objective: Improving data quality by identifying and correcting errors, inconsistencies, and duplicates.

Benefits: Enhanced data accuracy, improved decision-making, and increased operational efficiency.

Duplicate Data

Objective: Identifying and removing duplicate records to maintain a clean and accurate database.

Benefits: Reduced storage costs, improved data integrity, and enhanced user experience.

Data Integration

Objective: Integrating data from different sources by matching and linking related records.

Benefits: Seamless data integration, unified data view, and improved data analysis capabilities.

Data Silos

Objective: Breaking down data silos by integrating disparate datasets using fuzzy matching.

Benefits: Eliminated data silos, improved data accessibility, and enhanced collaboration across departments.

Deploying New Business Systems

Objective: Migrating data to new business systems by ensuring data consistency and accuracy and you're often deploying 'mismatching' data from multiple systems.

Benefits: Smooth system migration, reduced data migration errors, and minimized business disruptions.

Data Analysis

Objective: Enhancing data analysis by linking related records and providing a more comprehensive view of the data.

Benefits: Improved data insights, better trend identification, and enhanced predictive analytics capabilities.

Master Data Management

Objective: Managing master data by maintaining a single version of truth across the organization.

Benefits: Unified data governance, improved data quality, and streamlined business processes.

List Management

Objective: Managing marketing or customer lists by identifying and merging duplicate entries.

Benefits: Targeted marketing campaigns, improved customer segmentation, and enhanced customer engagement.

Customer 360 View

Objective: Creating a unified view of customers by linking related records from different data sources.

Benefits: Enhanced customer understanding, personalized marketing strategies, and improved customer satisfaction.

Vendor 360 View

Objective: Creating a comprehensive view of vendors by linking and consolidating vendor records.

Benefits: Improved vendor management, optimized procurement processes, and reduced supply chain risks.

Product 360 View

Objective: Creating a complete view of products by linking related records and categorizing products accurately.

Benefits: Efficient inventory management, targeted marketing, and improved product lifecycle management.

3. Different Fuzzy Match Algorithms

Soundex

Use Case: Phonetic matching
Pros: Good for names that sound similar
Cons: Slower and limited in handling typographical errors
Best Suited For: Company names, contact names

Levenshtein Distance

Use Case: String similarity
Pros: Flexible and can handle typographical errors
Cons: Computationally expensive for large datasets
Best Suited For: Text fields, product names

Jaro-Winkler

Use Case: String similarity with weight for common prefixes
Pros: Effective for short strings and similar names
Cons: Less effective for longer strings
Best Suited For: Contact names, addresses

Metaphone

Use Case: Phonetic matching
Pros: Handles variations in spelling and pronunciation
Cons: Limited to English language
Best Suited For: Contact names, product names

n-gram similarity

Use Case: Substring matching
Pros: Effective for identifying similar substrings
Cons: Sensitivity to substring length and order
Best Suited For: Text fields, descriptions

TF-IDF

Use Case: Text matching based on term frequency
Pros: Effective for text data and document similarity
Cons: Complex to implement and computationally intensive
Best Suited For: Text fields, documents

4. Business Case Examples for No-Code Fuzzy Match Solutions

Increasing Data Workloads

Challenge: With increasing data volumes, manual data matching becomes impractical, and most business people rely on third parties to prepare and process data.

Solution: No-code fuzzy match solutions simplify the process and improve efficiency, allowing organizations to handle larger data workloads without compromising accuracy.

Coding Challenges

Challenge: Coding fuzzy match algorithms from scratch or using libraries can be time-consuming, much less flexible, and error-prone.

Solution: No-code solutions provide a user-friendly interface for easy implementation, reducing the dependency on coding expertise and accelerating the deployment process. These solutions can be deployed very quickly and easily and provide a much higher level of configurability which is important for mismatching data.

Data Matching Challenges

Challenge: Most data doesn't match perfectly due to variations and errors.

Solution: No-code fuzzy match solutions provide flexibility and the configurability to handle these variations without any coding, ensuring accurate and reliable data matching across different domains. This reduces time to value, increases the number of good matches, and decreases the number of bad matches (false positives).

Ease of Use

Challenge: Business people often need to work with data without technical expertise and often need to work with data independently (without going to IT or data engineers).

Solution: No-code fuzzy match solutions are designed to be simple and intuitive, allowing business users to manage data effectively without requiring specialized technical skills.

Operational and Analytical Value

Benefits: No-code fuzzy match solutions not only enhance data quality but also provide valuable insights for operational and analytical purposes.

5. Data Matching Efficiency

As explained in his blog at Liliendahl on Data Quality, Henrik Liliendahl explains that there are 5 typical approaches to data matching:

1. Simple deterministic - efficiency 20%-50%

2. Synonyms / standardization - efficiency 30%-60%

3. Algorithms - efficiency 40%-70%

4. Combined traditional - efficiency 50%-80%

5. AI enabled - efficiency 10%-90%

Most business systems like CRM's, ERP's and others only use simple deterministic matching, which is why relational data is missed. Purpose-built solutions use one or some combination of all of these approaches, ranging from deterministic to AI enabled. Match Data Pro uses all of these approaches, in a simple-to-use and configurable user interface.

6. Fuzzy Matching and Entity Resolution Vendors

Choosing the right vendor for fuzzy matching and entity resolution solutions depends on your organization's technical expertise and specific requirements. Below are two categories of vendors: software built for IT or engineers and no-code software designed for business people. Each vendor is accompanied by details on pricing, technical skills needed, and specific requirements.

Software Built for IT or Engineers

1. Senzing

- **Website:** https://www.senzing.com
- **Description:** Senzing offers real-time AI for entity resolution, providing advanced fuzzy matching tools for data linking and deduplication.
- **Price Range:** $$$ (High)
- **Technical Skills Needed:** Advanced
- **Note:** Senzing provides sophisticated solutions tailored for organizations with advanced technical expertise in data management and AI.
- **SaaS/Monthly Subscriptions Available:** unknown

2. Talend

- **Website:** https://www.talend.com

- **Description:** Talend provides data integration and integrity solutions with robust fuzzy matching capabilities.
- **Price Range:** $$$ (High)
- **Technical Skills Needed:** Advanced
- **Note:** Talend offers comprehensive solutions suitable for organizations with advanced technical expertise in data integration and management.
- **SaaS/Monthly Subscriptions Available:** unknown

3. Informatica

- **Website:** https://www.informatica.com
- **Description:** Informatica offers comprehensive data management solutions, including advanced fuzzy matching and entity resolution capabilities.
- **Price Range:** $$$ (High)
- **Technical Skills Needed:** Advanced
- **Note:** Informatica provides enterprise-grade solutions ideal for large organizations with complex data management requirements.
- **SaaS/Monthly Subscriptions Available:** unknown

4. IBM

- **Website:** https://www.ibm.com
- **Description:** IBM offers a range of data management and analytics solutions, including powerful fuzzy matching algorithms.
- **Price Range:** $$$ (High)
- **Technical Skills Needed:** Advanced
- **Note:** IBM's solutions are designed for large-scale deployments and require advanced technical expertise for implementation and management.
- **SaaS/Monthly Subscriptions Available:** unknown

5. SAS

- **Website:** https://www.sas.com
- **Description:** SAS provides advanced analytics and data management solutions, including state-of-the-art fuzzy matching and entity resolution tools.
- **Price Range:** $$$ (High)
- **Technical Skills Needed:** Advanced

- **Note:** SAS offers comprehensive analytics solutions with sophisticated features, suitable for organizations with advanced technical capabilities.
- **SaaS/Monthly Subscriptions Available:** unknown

6. Ataccama

- **Website:** https://www.ataccama.com
- **Description:** Ataccama offers AI-powered data quality and master data management solutions, including sophisticated fuzzy matching algorithms.
- **Price Range:** $$$ (High)
- **Technical Skills Needed:** Advanced
- **Note:** Ataccama's solutions leverage AI technologies and require advanced technical skills for implementation and customization.
- **SaaS/Monthly Subscriptions Available:** unknown

7. Experian Data Quality

- **Website:** https://www.edq.com
- **Description:** Experian Data Quality offers comprehensive data quality management solutions, including robust fuzzy matching tools.
- **Price Range:** $$ (Moderate)
- **Technical Skills Needed:** Intermediate to Advanced
- **Note:** Experian Data Quality provides robust solutions suitable for organizations with varying levels of technical expertise.
- **SaaS/Monthly Subscriptions Available:** unknown

8. Innovative Systems

- **Website:** https://www.innovativesystems.com
- **Description:** Innovative Systems provides enterprise data management solutions, including powerful fuzzy matching and entity resolution capabilities.
- **Price Range:** $$$ (High)
- **Technical Skills Needed:** Advanced
- **Note:** Innovative Systems offers comprehensive solutions with advanced features, ideal for large organizations with advanced technical requirements.
- **SaaS/Monthly Subscriptions Available:** unknown

No-Code Software for Business People

1. Match Data Pro

- **Website:** https://www.matchdatapro.com
- **Description:** Match Data Pro offers intuitive fuzzy matching tools designed for business users without requiring coding skills.
- **Price Range:** $ (Ranges from Low to High)
- **Technical Skills Needed:** Beginner
- **Note:** Match Data Pro provides user-friendly solutions suitable for small businesses and organizations with limited technical expertise.
- **SaaS/Monthly Subscriptions Available:** Yes
- **Multilingual:** Yes
- **Multiuser:** Yes
- **OS Requirements:** Containerized
- **Built for Private Cloud:** Yes
- **Browser-based:** Yes

2. Winpure

- **Website:** https://www.winpure.com
- **Description:** Winpure offers easy-to-use data cleaning and fuzzy matching software designed for business users.
- **Price Range:** $ (Ranges from Low-High)
- **Technical Skills Needed:** Beginner
- **Note:** Winpure provides straightforward solutions with intuitive features, making it accessible for organizations with limited technical expertise.
- **SaaS/Monthly Subscriptions Available:** unknown
- **Multilingual:** Yes
- **Multiuser:** Single user/unknown
- **OS Requirements:** Windows/unknown
- **Built for Private Cloud:** unknown
- **Browser-based:** Windows/unknown

3. Data Ladder

- **Website:** https://www.dataladder.com
- **Description:** Data Ladder offers data quality and matching solutions, including intuitive fuzzy matching tools.
- **Price Range:** $$ (Ranges from Moderate to High)
- **Technical Skills Needed:** Beginner to Intermediate
- **Note:** Data Ladder provides user-friendly solutions with intuitive features, suitable for organizations with limited technical expertise.
- **SaaS/Monthly Subscriptions Available:** No
- **Multilingual:** Yes
- **Multiuser:** No
- **OS Requirements:** Windows/Linux
- **Built for Private Cloud:** No
- **Browser-based:** No

4. Melissa

- **Website:** https://www.melissa.com
- **Description:** Melissa provides global data quality and address verification solutions, including advanced fuzzy matching algorithms.
- **Price Range:** $$ (Ranges from Moderate to High)
- **Technical Skills Needed:** Beginner to Intermediate
- **Note:** Melissa offers straightforward solutions with advanced features, suitable for organizations looking for a balance between functionality and ease of use.
- **SaaS/Monthly Subscriptions Available:** Yes
- **Multilingual:** Yes
- **Multiuser:** Yes
- **OS Requirements:** unknown
- **Built for Private Cloud:** unknown
- **Browser-based:** Yes

5. Alteryx

- **Website:** https://www.alteryx.com

- **Description:** Alteryx provides a platform for data blending, analytics, and visualization, including user-friendly fuzzy matching tools.
- **Price Range:** $$$ (Ranges from Low to High)
- **Technical Skills Needed:** Beginner to Intermediate
- **Note:** Alteryx offers a comprehensive platform with intuitive features, suitable for organizations with varying technical skill levels.
- **SaaS/Monthly Subscriptions Available:** unknown
- **Multilingual:** Yes
- **Multiuser:** Yes
- **OS Requirements:** unknown
- **Built for Private Cloud:** unknown
- **Browser-based:** unknown

Conclusion

When selecting a vendor for fuzzy matching and entity resolution solutions, it's essential to consider factors such as pricing, technical skills needed, and specific requirements. Whether you're an IT professional looking for advanced solutions or a business person seeking user-friendly tools, there's a vendor that can meet your needs and help you achieve accurate and efficient data matching and entity resolution.

Related Keywords

Fuzzy matching, record matching, data matching, entity resolution, record linkage, merge purge, text matching, probabilistic matching, name matching, vendor matching, patient matching, householding, address matching, product matching, deduplication

Sources

[1] https://zipdo.co/statistics/data-management/

[2] https://liliendahl.com/2024/04/16/data-matching-efficiency/