

Are Duplicate Records Eroding Your Bottom Line?

*The Case for Fuzzy Matching to
Find Not-So-Obvious Duplicates*

Are Duplicate Records Eroding Your Bottom Line?

Fuzzy matching and deduplication help achieve a single view of customers for competitive advantage.

Introduction

Bad data is a challenge for all organizations. A recent TDWI study reported that organizations lose more than \$611 billion each year due to bad data.

And one major root cause of bad data is duplicate records. Conflicting data is corruptive to the integrity of databases and prevents organizations from gaining a single, accurate and organized view of enterprise data. Poor data quality that includes duplicate records restricts the implementation of mission-critical initiatives and creates the risk of poor business decisions by clouding critical customer, employee, and financial information.

Duplicate records come from many sources including acquisitions or mergers, legacy systems, data migrations and data entry errors. Regardless of the source, these issues quickly became a costly expense for your business in program inefficiencies, missed opportunities, and erroneous views of information.

There is, however, a way to turn this chaotic data into actionable data – by identifying and eliminating duplicate records. Deduplication of data through the use of merge/purge solutions is an important component of the data cleansing process. But identifying duplicate records has its own set of challenges.

Using Fuzzy Matching to Identify Not-So-Obvious Duplicates

For exact duplicates where records are identical, deduplication is a relatively simple process. However, most data contains “non-exact matching” duplicate records that are difficult to identify.

For example, a ‘Beth Smith’ at ‘United Data Machines’ could also be recorded in the same or different database as ‘Smithe, Elizabeth’ at ‘UDM.’ In reality, Beth Smith and Elizabeth Smithe are one and the same, but your organization would think they are two different contacts.

To find hard-to-spot duplicate records requires fuzzy matching logic.

What is Fuzzy Matching?

Fuzzy matching is an advanced mathematical process that determines the similarities between data sets, information, and facts – where the outcome is neither true nor false, or 100 percent certain, hence the word, “fuzzy.” The process compares any data type of any length and from any place in a field to find non-exact matches.

According to data mining consulting firm Two Crows Consulting, fuzzy matching or fuzzy logic is “applied to fuzzy sets where membership in a fuzzy set is a probability, not necessarily 0 or 1... Fuzzy logic needs to be able to manipulate degrees of maybe, in addition to true and false.”

For every piece of data examined, the fuzzy matching process will give a probability score to determine the accuracy of the match. For example, 'Tomas Jones' might get a 90 percent score of similarity, while 'Tom Jones' might receive a 75 percent score, as compared to the actual name of Thomas Jones.

To demonstrate how duplicate records are identified through fuzzy matching, here is a sample list of prospective customers. As you can see, there are duplicate records due to misspelling or typos. Customer # 11 and Customer #111, and Customer #1111 are most likely the same person.

Still, there are other methods – such as the use of more advanced fuzzy matching algorithms – that can identify whether Customer #11, Tomas Jones, Customer #111, Tom Jones, and Customer #1111, Thhomas Jones are indeed the same person.

Example #1

Cust_Id	Name_Input	Last_Name
11	Tomas	Jones
22	Peter	Jackson
33	Theresa	Smith
333	Therese	Smith
111	Tom	Jones
222	Pete	Jackson
55	Johnathan	O'Conner
555	Johnatiag	O'Conner
1111	Thhomas	Jones
2222	Petet	Jackson
3333	Theses	Smith
5555	Jonathan	O'Conner

Different Fuzzy Matching Algorithms

There are several different algorithms that a deduplication or merge/purge program can employ in the fuzzy matching process. These algorithms perform more effective searches of full names in databases.

Phonetic Matching

Utilizing the phonetic algorithm for fuzzy matching – also known as the “double metaphone” method – detects “alike-sounding” relationships between words. For example, phonetic matching will identify a sound-alike relationship between 'Folsom' and 'Folsum,' and no phonetic connection between 'George' and 'Mark.'

Basically, phonetic matching allows you to perform approximate searches, instead of just 'exact' matches – thus enabling organizations to find variations of a first name or last name.

According to U.K.-based data services firm KnockYourSocksOff.com, double metaphone/phonetic matching can compute primary and secondary encoding for a given word or name to determine the most likely pronunciation, and alternative pronunciation. The process calculates that some words can have more than one pronunciation.

Phonetic matching is best used for detecting duplicates with proper names, than it is with larger fields of generic text. Phonetic matching replaced the now obsolete Soundex matching algorithm.

N-gram or Q-gram-based Algorithms

The linear n-gram or q-gram-based algorithms models are primarily used in statistical natural language processing. An n-gram is a subsequence of n items from a given sequence – which can be phonemes, syllables, letters, words, or base pairs, as defined by Wikipedia.

Monograms, bigrams, and trigrams are examples of a q-gram. According to The Code Project – a development resource – q-gram algorithms aren't strictly phonetic matching in that they do not operate based on comparison of the phonetic characteristics of words. Instead, q-grams can be thought to compute the "distance," or amount of difference between two words. Utilizing the n-gram or q-gram algorithm method is highly favorable, as it can match misspelled or mutated words, even if they are determined to be "phonetically disparate."

For example, the word 'Nelson' has the following q-grams: **Ne el ls so on.**

To compare the difference in matching, the word 'Neilsen' is broken down into this q-gram: **Ne ei ill s se en.** Because the words were similar sounding (even though they were spelled differently), they were identified as a match.

Jaro-Winkler Algorithm

The Jaro-Winkler distance is a measure of similarity between two strings. It is mainly used in the area of record linkage for duplicate detection. The higher the Jaro-Winkler distance for two strings, the more similar the strings are. The Jaro-Winkler distance metric is best suited for short strings such as person or proper names. The score is normalized such that 0 equates to no similarity and 1 is an exact match.

More Fuzzy Algorithms

Here are other fuzzy algorithms that can be used in the matching process:

- **Containment** — Matches when one record's component is contained in another record. For example, "Smith" is contained in "Smithfield."
- **Frequency** — Matches the characters in one record's component to the characters in another without any regard to the sequence. For example "abcdef" would match "badcfe."
- **Fast Near** — A typographical matching algorithm. It works best in matching words that don't match because of a few typographical errors. Exactly how many errors is specified on a scale from 1 to 4 (1 being the tightest). The Fast Near algorithm is a faster approximation of the Accurate Near algorithm.
- **Accurate Near** — This is a typographical matching algorithm. The Accurate Near algorithm produces better results than the Fast Near algorithm, but is slower.

The Cost of Duplicate Records

Did you know that an estimated 10 percent of the names and addresses in your average mailing list are duplicate records? That means that if your mailing list contains 10,000 records, with production and postage costs averaging 83 cents per piece – your total mailing cost would be \$8,300. So, if 10 percent of your list is made up of duplicate records, you are wasting \$830 every time you mail. That's a big blow to your bottom line.

Costly:

- Direct marketing communications are repeated unnecessarily
- Product shipments could be sent to wrong addresses
- Sales reporting may be inaccurate
- Wasted postage, paper and printing

Embarrassing:

- Duplicate communications give an impression of sloppiness
- No accurate, single view of customer sales and communication history

- **Frequency Near** — Similar to Frequency matching except that the algorithm lets you specify how many characters may be different between components.
- **Vowels Only** — Only vowels will be compared. Consonants will be removed from the search criteria.
- **Consonants Only** — Only consonants will be compared. Vowels will be removed from the search criteria.
- **Alphas Only** — Only alphabetic characters will be compared.
- **Numerics Only** — Only numeric characters will be compared. Decimals and signs are considered numeric.

Real Case Scenario:

A Las Vegas hotel casino makes a reservation for a customer under the name Johnathan Smith. Believing Johnathan Smith is a new customer, the hotel casino's marketing department sends him an email with a free night's stay promotion as part of its new customer incentive program. The marketing department was unaware that Johnathan Smith is already an existing customer, listed in the hotel casino's gaming department as John Smith, a VIP client. Based on his previous customer history, John Smith is eligible for several upgrades as part of their customer loyalty program that would make his stay more enjoyable, and help guarantee the hotel casino continues to enjoy future patronage from a high-value customer.

The Problem:

Duplicate records in the hotel casino's database obscure an accurate view of customers and their buying behavior, preventing efficient and effective marketing efforts to help retain and gain additional business or build loyalty with high-value customers.

The Solution:

The hotel casino used a merge/purge program to first find all non-exact matching records in the database. Then, based on their business needs, developed a merging process that would find the first record created for that customer and credit that "creation date" as the Member Since date for the customer loyalty program. The hotel casino also needed to develop a process that would tabulate all of the customer loyalty points earned in each of the duplicate records and add those to the master file. Now the hotel casino has an accurate view of high-value customers – it knows when they became a customer and how much each spent on hotel reservations, food, and other amenities that accrued which can be tabulated as customer loyalty points. Armed with this information, the hotel casino's marketing department can now, with confidence, push out marketing campaigns that are relevant and effective in helping increase patron loyalty of high-value VIP customers, efficient and effective marketing efforts to help retain and gain additional business, or build loyalty with high-value customers.

Different Ways to Implement Deduplication

Some vendors allow customers to customize exactly how to implement deduplication technology into their operations – thus enabling the flexibility to integrate the process at different points of the business process.

Here is a three-point checklist to follow when customizing your deduplication process:

1. Standardize your data. – Accurate, consistent, high-quality data is the foundation of a database, so the need to implement a data standardization process to ensure the integrity and validity of your data – is critical. An organization can't build a data warehouse, integrate or migrate data, or get a complete and accurate view of their customers without first standardizing data. There are data quality solutions in the market that can standardize, verify, and validate the contact information in your database, some even in real-time at point-of-entry, or in batch mode.

2. Determine your business needs. – Based on your specific business needs, you can compare your records at once, which is ideal for batch merge/purge suppressing existing data; or you can compare each record as they come in and against a database of already processed records – ideal for real-time data entry. Another method is called “hybrid deduping” that allows businesses to customize how records are processed and stored.

3. Monitor your data. – Databases that contain contact data are never static because information is constantly changing as customers move, change companies, die, or divorce. And, new or possibly inconsistent data is coming in all the time from call centers, Web forms and data entry by various departments. Therefore, it is important to add a monitoring package into your operations – this will provide real-time data monitoring in an automated process to immediately recognize and correct issues before the quality of data declines. This approach also helps organizations enforce data governance and compliance measures.

Conclusion

The true value of any database is determined by one fundamental component: the quality of the data. Without data that is reliable, accurate, and updated, organizations can't deliver trusted customer, product, and other vital data throughout the enterprise. One hidden culprit of bad data is duplicate records, which results in waste, operational inefficiencies, lost revenue, less effective sales and marketing, and poor customer service.

Consequently, the deduplication of data is one of the most critical components in the data cleansing process. But before adopting these deduplication methods, an organization must also integrate a data standardization process first and foremost. Identifying duplicate records is challenging in and of itself. The biggest roadblock to identifying duplicates lies in detecting non-exact matching duplicate records – the data that appears to be multiple sets of information, but are actually duplicate records. There are ways to overcome these challenges – the most successful method is by employing fuzzy matching algorithms as part of a merge/purge process. Detecting the most duplicate records helps streamline databases, improve marketing efficiency, and achieve a unified, accurate view of the customer.

When looking for a solution for data matching and deduplication, look for a program that can perform both exact matching and the latest fuzzy matching techniques to match across multiple columns or, across multiple data sources, to identify duplicates and manage your master data with incremental comparisons.

About Melissa Data Corp.

Melissa Data (www.melissadata.com) is a leading provider of data quality, data integration and data enrichment solutions. The company recently released its new multiplatform MatchUp Object, an API toolkit for finding and preventing duplicate records. MatchUp is also available as a desktop GUI. Melissa Data is an active DMA member, as well as a member of the DMA's List & Database Council. For over 25 years, Melissa Data has been a leading provider of data quality solutions with emphasis on U.S., Canadian, and international address and phone verification, and postal automation software. Free trial software is available by visiting www.MelissaData.com or by calling 1-800-MELISSA (800-635-4772).

© 2011 Melissa Data. All rights reserved.