Identity Resolution for Data Quality and Master Data Management

A White Paper by David Loshin

WHITE PAPER

Table of Contents

The Challenge of Unique Identity

The drive for data unity creates many opportunities for data consolidation. Customer data integration projects, master product catalogs, security master projects and enterprise master patient indexes are all examples of technology-driven projects intended to reduce multiple data sets containing similar information into a single view. The hope is that a unified data asset will lead to improved business processes. The success of this data integration process hinges on determining when different data instances in the same (or across multiple) data sets refer to the same real-world entity. Searching data sets for matching records is the focus of the data consolidation process.

The two most interesting challenges for data integration are basically two sides of the same coin: sometimes the challenge is determining when two records refer to the same real-world object; other times, it's about knowing for certain they do not refer to the same real-world object. Yet without being able to make that clear connection or distinction, it would be difficult – if not impossible – to identify potential duplicate records within and across data sets.

The method used to find these connections is typically referred to as identity resolution. From a technology perspective, identity resolution is a collection of algorithms used to parse, standardize, normalize and then compare data values. Identity resolution can establish that two records refer to the same entity or can determine that they don't. By feeding the record set into the identity resolution process, we can determine, for example, that each of these records contains a reference to a unique entity. Beyond that, we can use data extracted from all of the records to create a high-quality representation of each entity type. This process is used to resolve different entity representations and to determine if they all refer to the same real-world entity.

The techniques used in this process are critical for any business applications that rely on customer or product data integration as part of a master data management (MDM) and data quality initiative. In this paper, we explore the root cause of the dual challenge of identity resolution, examine how parsing and standardization contribute to the process, then review different ways that similarity scoring and approximate matching algorithms can help determine and resolve identical entities despite variations.

Entities and Their Attributes

A common theme in any data integration effort is the identification of a unique entity. But what is an entity, and how does it relate to the data integration process? For the purposes of identity resolution, an entity is an instance of a core data concept that is used in transactional, operational and analytical applications. Standard examples include "party," "customer," "product," "part," etc. A core data concept is usually based on various data models and is intended to reflect that concept's core characteristics. In a perfect world, each entity is unique within the data set and can be differentiated from any other instance stored within the data set.

However, the proliferation of databases and applications creates opportunities for redundancy and variation due to:

- Similar structures. Often the same underlying concepts interact in different roles, such as individuals who are manifested as employees, customers or beneficiaries. In this case, different relational structures may be used to represent the concept of an individual in any of the roles, and each model may carry different data attributes – or even the same attributes but with different data element names and data types.
- Similar content. Data values vary, especially in semi-structured attributes such as individual names or product names. In this situation, different values may be used in different records, even if they represent the same real-world entity.

One common aspect of any entity is its name. People, products, documents and any real-world object all have some kind of name used for reference. Yet what is the difference between an entity and the names that are used to refer to that entity? An object's name is just a collection of character symbols, usually assigned by other individuals and used as a tag to refer to that object. There is nothing intrinsically defining about an individual's name or a product's name, nor is it particularly unique or distinguishing. In fact, one person might be known by a number of names (Jon Smith, Jonathan Smith, Dr. J.M. Smith), each meaningful within a specific context.

As a simple example, consider the names used to refer to baseball legend Ty Cobb; aside from his given name, he was also referred to by a nickname, "The Georgia Peach." On the other hand, that same nickname might also be used to refer to a completely different entity in some different context. For example, "Georgia Peach" can also refer to a variety of peach that grows in the state of Georgia. In some cases, the data values are not particularly useful when it comes to differentiation. A quick scan of an online phone directory will yield hundreds, if not thousands, of individuals sharing the name "John Smith." Yet even individuals with uncommon names such as "David Loshin" still might not be distinguishable – a search at an online book store will show that there are two authors with that name.

As a second example, consider the list of bank entities shown in Figure 1; some of these records refer to individuals, while others refer to "account names" associated with different kinds of financial products or accounts. But at what point do we differentiate between the concept of an individual vs. the concept of an individual acting in a particular role with respect to another entity?

BARBARA GOLDFINGER LIVING TST	DTD 4/5/00 BARBARA GOLDFINGER	STEPHEN GOLDFINGER TRUSTEES
BARBARA M GOLDFINGER FAM TST	DTD 4/5/00 STEPHEN GOLDFINGER	& EDWARD G GOLDFINGER TSTEES
BARBARA M GOLDFINGER MASS QTIP	TST DTD 4/5/00 STEPHEN E	& EDWARD G GOLDFINGER TTEES
DAVID GOLDFINGER	6 CHANDLER STREET	LEXINGTON, MA 02420
EDWARD GOLDFINGER	950 FOUNTAIN STREET	ANN ARBOR, MI 48103
HENRY GOLDFINGER	TTEE 3/10/83 HENRY GOLDFINGER	LIVING TRUST
MATILDA T GOLDFINGER	TTEE 3/10/83 M T GOLDFINGER	LIVING TRUST
MICHAEL GOLDFINGER	11 CRESCENT HILL AVE	LEXINGTON, MA 02420
PETER GOLDFINGER	7506 HAMPTON AVE	LOS ANGELES, CA 90046
STEPHEN GOLDFINGER	THE GOLDFINGER FAMILY ACCOUNT	33 BIRCH HILL ROAD

Figure 1: Selected records from a public data set.

In this example, we have an entity for an individual, “Barbara Goldfinger,” and an entity for an account, “Barbara Goldfinger Living TST DTD 4/5/00.” Are these the same entity? Actually, no. The account entity is one in which the named individual acts in the role of trustee. In fact, a careful review of the 10 records in Figure 1 reveals 27 different entities, which can be shown in Figure 2.

BARBARA GOLDFINGER
BARBARA M GOLDFINGER
BARBARA GOLDFINGER LIVING TST DTD 4/5/00
BARBARA GOLDFINGER LIVING FAM TST DTD 4/5/00
BARBARA M GOLDFINGER MASS QTIP TST DTD 4/5/00
BARBARA GOLDFINGER TRUSTEE
BARBARA GOLDFINGER TSTEE
STEPHEN GOLDFINGER
STEPHEN GOLDFINGER TRUSTEE
STEPHEN GOLDFINGER TSTEE
STEPHEN E GOLDFINGER
STEPHEN E GOLDFINGER TTEE
EDWARD GOLDFINGER
EDWARD G GOLDFINGER
EDWARD G GOLDFINGER TSTEE
EDWARD G GOLDFINGER TTEE
DAVID GOLDFINGER
HENRY GOLDFINGER
HENRY GOLDFINGER TTEE
MATILDA T GOLDFINGER

MATILDA T GOLDFINGER TTEE
MICHAEL GOLDFINGER
PETER GOLDFINGER
THE GOLDFINGER FAMILY
THE GOLDFINGER FAMILY ACCOUNT
3/10/83 HENRY GOLDFINGER LIVING TRUST
M T GOLDFINGER

Figure 2: Out of 10 records, 27 entities are identified.

These examples highlight that, despite the arbitrariness of an object's name, it often carries additional descriptive content that could be used to describe more than one entity, either in different conceptual domains or even (in some instances) within the same domain. But if different objects have the same or similar characteristics, how do you differentiate them? More to the point: what are the characteristics of any entity set that can be used for unique identification, and consequently, for record matching and consolidation?

Identifying Attributes

The need for unique identification is inextricably linked to the success of any data consolidation project, but automating the matching process remains challenging, especially in the presence of semi-structured or unstructured data values, data errors, misspellings or words that are out of order. The existence of variable meanings of values appearing in free-formed text attributes also raises several questions:

- How can automated algorithms parse and organize values and determine which entities are represented?
- How many times and in how many different ways can variations occur?
- How can the identities be distinguished or resolved?

Automated identity resolution requires techniques for approximate matching that compare a variety of entity characteristics in a search for similarity.

Unique identification relies on comparing a combination of intrinsic attributes (eye color, for example) and assigned attribute values (name) to distinguish one entity from another. Name alone may not be enough, nor name plus other intrinsic attributes, but there is a set of "identifying attributes" whose combined values uniquely define an entity. For any collection of entity records, there should be some set of data attributes that can be used for unique identification; otherwise, there can be exact duplicates in the data set, which would violate the expectation that each entity is represented once and only once in the data set.

Determining which data elements can be used as identifying attributes becomes a critical task, whether the objective is data cleansing, MDM or other data consolidation. There are a variety of ways that these attributes can be included in an entity model with various data attribute names, data types, sizes and structures. Sometimes many data attributes are collected into a very large table, while other (perhaps more normalized) models have a relational structure allowing for a more flexible connectivity with different types of characteristics.

Attributes should be evaluated based on how well their constituent semantics and values contribute to addressing the dual challenge – enough information to distinguish two records representing different entities and enough to link two records representing the same entity. Some qualitative dimensions for evaluation include:

- Inherence – the degree to which the attribute is intrinsic to the entity. Examples include engineering specifications of a product, such as the "head diameter," "shank diameter," or "threading type" of a screw.
- Structural stability – the degree to which the attribute's structure is subject to variance. Attributes relying on a well-defined value domain (such as a salary range of $10,000 to $1,000,000) have a high degree of structural stability. Attributes like dates, telephone numbers and individual names can appear in a variety of patterns or formats and have a medium degree of structural stability. Free-form text values have a low degree of structural stability.
- Value stability – the degree to which the attribute's value changes and the change frequency. An example of a stable value is an individual's eye color.
- Domain cardinality – this looks at the size of the domain of a value. Attributes that use a domain with many possible values are more likely to be used for differentiation than those using a domain with a small number of values. For example, a birth date domain may have a limited set of 366 values if the birth year is excluded.
- Completeness – attributes that are missing data are less likely to contribute significantly to differentiation.
- Accuracy – attributes with a high degree of trust may be more reliable for similarity comparisons.

Even though the criteria may be different depending on the entity type and data quality, establishing an evaluation process based on the above criteria should simplify the selection of identifying attributes and lead to more effective choices for automated record matching and similarity scoring that makes identity resolution possible.

Approximate Matching and Similarity Scoring

When you think about it, automated identity resolution for record linkage is based purely on the fact that errors creep into data sets and prevent matching algorithms from working. Therefore, we use more complex methods to determine when there are enough similar data values between two records to reasonably presume that the records refer to the same entity.

The approximate matching process uses a number of strategies for similarity scoring that are intended to measure the conceptual distance between two sets of values. The closer the two sets of values, the more similar those two records are to each other. For each data type or data domain, we assign a similarity function. For each set of data attributes, a weight may be factored when computing an overall similarity score.

For example, given a set of data records with name, address, telephone number and birth date, we can configure a similarity function that is composed of the weighted similarity functions associated with each identifying attribute. Because "name" may contribute more to unique identification than "birth date," we'd assign a greater weight to name. Telephone numbers may contribute even more to unique identity, suggesting that telephone number be weighted even greater than name. Each data attribute's weight is often based on the measures associated with the criteria used to select the identifying attributes in the first place.

Each data type is subjected to its own similarity scoring method. Integer values may be scored based on a simple distance function – the closer the values are, the higher the score. Although scoring the distance between two numbers is straightforward, a comparison of string-based attributes (such as names, street addresses, descriptions or dates) is more complex, requiring qualitative measures to calculate distance. These measures often look at perceived similarity between the character strings and rely on approximate matching techniques, such as:

- Parsing and standardization – in addition to their use for data cleansing, these techniques can reduce the search space by linking entities. For example, standardization rules can be used to map from known names to a normal form that can then be used for similarity analysis. This technique is not limited to names and can be used for other kinds of data: product types, business words, addresses, industry jargon and transaction types, to name a few.
- Abbreviation expansion – similar to standardization, abbreviation expansion is a rule-oriented process that maps shortened forms to expanded forms to support the similarity analysis process. Abbreviations come in different forms. One type of abbreviation shortens each word in a set to a smaller form, where the abbreviation consists of a prefix of the original data value. Examples include "INC" for incorporated, "CORP" for corporation, "ST" for street. Other abbreviations shorten the word by eliminating vowels or by contracting the letters to phonetics, such as "INTL," or "INTRNTL" for international, "PRGRM" for program, "MGR" for manager, etc. Additionally, there are acronyms such as "RFP" for request for proposal that are formed from the initial letter of each word.

- Edit distance – this is the minimum number of basic edit operations required to transform one string to the other. There are three basic edit operations: insertion (where an extra character is inserted into the string), deletion (where a character has been removed from the string), and transposition (in which two characters are reversed in their sequence). For example, the edit distance between the strings "INTERMURAL" and "INTRAMURAL" is 3, because to change the first string to the second, we would transpose the "ER" into "RE," then delete the "E" followed by an insertion of an "A."
- Phonetic comparison – this technique considers how similar two strings sound. Examples include Soundex, NYSIIS and Metaphone, all of which encode character strings based on mapping sets of similar-sounding consonants and vowels into a standard format. The expectation is that similar-sounding names would be encoded as the same or similar phonetic codes, which might then be subjected to other approximate matching techniques.
- N-gramming – this method considers any string composed of substrings, which are grouped in discrete "chunks" moving from left to right (for Western languages). Each string is broken into a set of chunks of size n by sliding a window of size n across the word, and grabbing the n-sized string chunk at each step. The chunk size is determined by the n in the n-gram. For example, when bigramming, the name "DAVID" is broken up into the four two-character strings "DA," "AV," "VI," and "ID." If two strings match exactly, they will share all the same n-grams, but if two strings are only slightly different, they will still share a large number of the same n-grams. This similarity measure between two strings compares the number of n-grams the two strings share.

This is a sampling of some of the matching and sampling techniques. These methods and other algorithms can be adjusted and improved through the incorporation of business rules, statistical analysis and predictive assessments that can more accurately (and potentially dynamically) adjust weighting factors to provide an accurate and trustworthy similarity score.

False Positives, False Negatives and Thresholding

These techniques all contribute to a numeric score of distance between two values, and those scores can be scaled and weighted to provide a single similarity score. The similarity score is used for identity resolution to determine when two records match. If there is a high degree of similarity, there is a greater likelihood that the records refer to the same entity. If there is a low degree of similarity, there is greater likelihood that the records refer to different entities.

The perception of "high" and "low" translate into a defined threshold for matching, and any score above that threshold indicates a match. Yet, if you recall our dual challenge of identity resolution, we must beware of the possibilities of two types of failures:

1. False positives, in which two records that do not represent the same entity are determined to be a match.

2. False negatives, in which two records that represent the same entity are not matched.

False positives occur when the match threshold is set too low, while false negatives happen when the match threshold is set too high. A more effective approach is to provide two thresholds: a match threshold and a no-match threshold. When the similarity score is above the match threshold, the process automatically deems the records to represent the same entity, and when the similarity score is below the no-match threshold, the records are deemed to represent different entities.

Scores that fall between the two thresholds require special attention for two reasons. First, because the identity resolution process was unable to discretely provide an answer, there is a need for manual review. In this situation, a subject-matter expert will look at the two records and decide whether they match or not. Second, by evaluating any patterns or commonalities in those record pairs selected for manual review, the similarity scoring algorithms can be tweaked to improve their precision and, consequently, improve matching. Iterative refinement of the similarity scoring algorithms will ultimately help to reduce the area between the match and the no-match thresholds. This fine-tuning will reduce the number of questionable records pairs – and substantially decrease the need for manual intervention.

Probabilistic vs. Deterministic – Does It Really Matter?

Another aspect of similarity scoring incorporates statistics as input to the determination of the weighting factors. While a deterministic approach relies on defined business rules that determine when a pair of records will match, a probabilistic approach incorporates some likelihood (usually expressed as a percentage) that two records will match. Informally, probabilistic algorithms consider frequency analysis of value sets associated with the identifying attributes, as well as looking at dependent variables that might affect scoring precision.

Recalling an earlier example, the fact that “John Smith” is a very common name means that it is less likely that two records associated with the name “John Smith” are going to be a match. On the other hand, two records associated with the same very uncommon name have a much higher probability of referring to the same individual.

The question often arises: which approach is better? The question suggests that one approach is objectively better than the other. In reality, the effectiveness of an algorithmic approach must be measured within the context of how well it helps achieve the intended business objectives. Some considerations include:

- Number of records.
- Required matching precision.
- Number of identifying attributes.

- Variation in identifying attribute values.
- Risk tolerance/business impacts of false positives.
- Risk tolerance/business impacts of false negatives.
- Performance.
- Traceability.
- Adaptability to changes over time.

Ultimately, deciding which approach is better depends on whether one choice significantly affects the way identity resolution meets defined business requirements. However, for many business applications, either approach is more than sufficient to satisfy the business needs.

Summary

The need to link and consolidate entity information with a high level of confidence depends on comparing identifying data within a pair of records to determine similarity between that pair. Identity resolution employs techniques for measuring the degree of similarity between any two records and is often based on weighted approximate matching between a set of attribute values between the two records.

A process to analyze the suitability of entity data elements as candidate-identifying attributes must accompany the selection of an identity resolution tool. This assessment must take a number of factors into consideration, especially when observing how well the attribute selection helps meet the dual challenge associated with unique identification and entity differentiation for record matching.

By applying approximate matching techniques to sets of identifying attributes, identity resolution can be used to recognize slight variations that suggest that different records are connected, where values may be cleansed or where enough differences exist between the data to suggest that the two records represent different entities. Identity resolution is a critical component of most data quality, MDM and business intelligence applications. Achieving customer centricity or a comprehensive product catalog depends on resolving identities for all records that carry information about each unique entity and then creating a unified view.

David Loshin, President of Knowledge Integrity Inc., is a recognized thought leader and expert consultant in the areas of data quality, master data management and business intelligence. Loshin is a prolific author regarding data management best practices and has written numerous books, white papers and Web seminars on a variety of data management best practices. His book *Business Intelligence: The Savvy Manager's Guide* has been hailed as a resource allowing readers to "gain an understanding of business intelligence, business management disciplines, data warehousing and how all of the pieces work together." His book *Master Data Management* has been endorsed by data management industry leaders, and his valuable MDM insights can be reviewed at mdmbook.com. Loshin is also the author of the recent book *The Practitioner's Guide to Data Quality Improvement.* He can be reached at loshin@knowledge-integrity.com.

About SAS

SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions, SAS helps customers at more than 65,000 sites improve performance and deliver value by making better decisions faster. Since 1976 SAS has been giving customers around the world THE POWER TO KNOW®.

SAS Institute Inc. World Headquarters+1 919 677 8000
To contact your local SAS office, please visit: **sas.com/offices**