Probabilistic Versus Deterministic Data Matching: Making an Accurate Decision

Information Management Special Reports, January 2007

Scott Schumacher

Consider the following scenario: One night paramedics wheel a man named John K. Spenser into a hospital emergency room where admitting staff work to determine his medical history. Because the facility had not invested in an effective identity matching solution, a query of the hospital database and area medical centers does not correctly identify the patient as John K. Spencer (different spelling). To further complicate matters, because there wasn't a match, the medical staff was not aware of an important medical detail that could mean the difference between life and death: Mr. Spencer has a history of anaphylaxis, a potentially fatal allergic reaction to Penicillin and clot-busting drugs used after heart attacks.

While fictional, the previous example illustrates a potential risk of relying on inadequate data matching technologies for identification purposes. Though such technologies have improved substantially in recent years, many businesses across all industries still rely on inaccurate matching mechanisms, making their customer data integration (CDI) systems prone to serious error. In the case of healthcare, using the wrong technologies can hinder and delay treatment efforts and the ability to comply with state and federal privacy laws. Inadequate data matching methods also jeopardize companies, governments and organizations that must rapidly determine the identities of individuals. For example, the quality of the technology can mean the difference between preventing a terrorist from boarding a jet airliner and allowing one to pass undetected.

In addition, inaccurate data matching can cost businesses thousands or, in some cases, millions of dollars.

- A financial institution could be fined for dealing with a prohibited foreign company because their matching logic failed to recognize that one of their customers was on a government watch list.
- An insurance company with inaccurate matching systems could fail to identify a supposedly new customer as an existing customer with a delinquent account.
- A retail firm with systems that lack real-time accurate identification could lose a customer by not recognizing them when they place an order over the Web instead of their usual method of ordering by phone.

Therefore, it is essential for companies to take steps to ensure their CDI system leverages the most effective identity matching solution.

Choosing a Matching System

Though accuracy has long been viewed as the cornerstone of any successful CDI installation, deciding which method to use to ensure precise automated data matching can be difficult. In the CDI industry, "matching" refers to the process of determining when two records belong to the same customer.

Because different industries require different degrees of accuracy and because the complexity and diversity of data sources vary by company, businesses need to decide which data matching strategy best suits their needs prior to implementation. Each method has its strengths and weaknesses, so businesses should not be concerned about which is better in an absolute sense, but rather what the practical application will be and their tolerance for errors. Unfortunately, there is no perfect system, so understanding the possible matching errors and their frequency is key in finding the right solution to meet an organization's needs. In CDI lingo, inaccuracies are expressed as " false positives" and "false negatives." False positives occur when the system mistakenly links records that should not be matched (mismatches); false negatives result when the system fails to link two records that should have been matched (missed matches). However, these inaccuracies can vary greatly depending on the type of matching method being used.

Understanding Matching Methods

In today's CDI industry there are essentially two methods available for matching and retrieving data in CDI installations: probabilistic and deterministic.

Deterministic Matching

Deterministic matching systems use a combination of algorithms and business rules to determine when two or more records match (the rule "determines" the result). In a deterministic matching system, for example, one rule might instruct the system to match two

records with different names if the Social Security number and address fields coincide. Algorithms catch simple common errors such as typos, phonetic variations and transpositions. The result is an either/or outcome: Either records match the requirements of the business rule or they don't.

Deterministic matching systems have a relatively lower degree of accuracy compared to probability matching. Such systems are best suited for applications where the number of records is relatively small (less than two million), there are few data attributes and there is no great consequence of error. One such application could be mailing list processing. If the system matches a name to an incorrect address, the mailing would be sent to the wrong person, resulting in the sending company wasting postage costs.

Deterministic systems do allow organizations to better leverage their in-house IT staff for system implementation and to develop matching rules. When the number of data attributes and rules required are small, this can make implementation times shorter and less expensive. However, the more attributes involved and the larger the data sets, the more complex the rules-based matching routines become. This means implementation can involve many man hours of development and testing time and longer deployment times than probabilistic systems. Deterministic approaches do not have speed advantages over probabilistic methods, which now have the capability to perform lookups in real time.

In addition, deterministic systems lack scalability. When databases grow beyond a few hundred thousand records companies with deterministic matching systems typically require expensive customization and business rule revision. If an attribute is added to a data set, this doubles the number rules the system requires, which can be very labor-intensive and impact system scalability and performance. Both examples push the maintenance costs and total cost of ownership of deterministic systems far higher than that of a probabilistic matching solution.

Probabilistic Matching

Probabilistic matching uses likelihood ratio theory to assign comparison outcomes to the correct, or more likely decision. This method leverages statistical theory and data analysis and, thus, can establish more accurate links between records with more complex typographical errors and error patterns than deterministic systems.

Typically, probabilistic systems assign a percentage (such as 75 percent) indicating the probability of a match. Because these systems pinpoint variation and nuances to a much finer degree than a deterministic approach, they are better suited for businesses that have complex data systems with multiple databases. Due to the size of these data systems, the potential for duplicates, human error, and discrepancies is far greater, making a system designed to determine links between records with complex error patterns much more effective.

In a probabilistic matching system, algorithms weigh frequency and uniqueness of data. Regional differences are factored into the equation, enabling a match on the name Jose Rodriguez, for example, to score differently in Los Angeles than it would in Saint Louis. Probabilistic systems check all possible name alternatives and consider variables such as nicknames, phonetics (Gerald versus Jerold), transposed last and first names, and use of initials (Chuck L. Jones versus Chuck Lawrence Jones or C. Lawrence Jones). If programmed, probabilistic algorithms match against international languages and dialects.

Probabilistic systems adapt to the data to which they are being applied and do not require much manual tuning to implement and maintain. Probabilistic algorithms can also easily accommodate a growing number of files and databases with no sacrifice in speed or accuracy.

For situations where data set sizes and numbers of attributes are large, and high levels of accuracy and low total cost of ownership are important, organizations should select a probabilistic system. When data sets are smaller, have fewer attributes and accuracy doesn't cost an organization much in terms of risks or consequences, then a deterministic approach may be preferable.

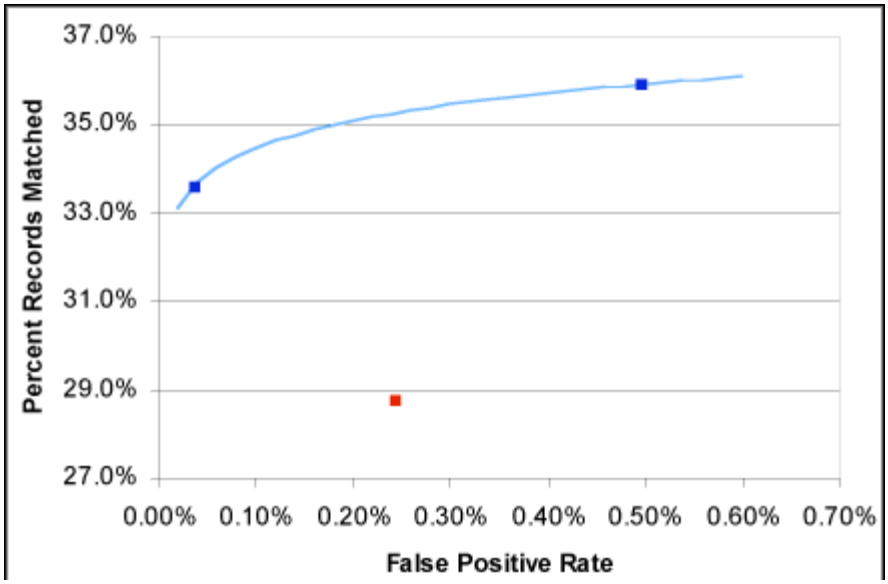

Figure 1: Comparison of Deterministic versus Probabilistic Matching

Figure 1 shows a real-world comparison of data from a deterministic and a probabilistic matching system. In this experiment, two matching systems were used to match 3 million new transaction records against a customer master containing 16 million records. The results are shown on an ROC (receiver operating characteristic) curve, an analysis method borrowed from radar detection. The figure illustrates on a single chart the false positive and false negative matching performance of both a deterministic and probabilistic system.

The deterministic method is a single point because deterministic systems do not allow optimization for false positives or false negatives - they are either/or decisions. System designers do not have control over what they want actual error rates to be. In addition, the single deterministic point is significantly lower and to the right of the probabilistic curve, which means that performance is worse for both false positives and false negatives than with the probabilistic system.

Unlike deterministic systems, probabilistic data is mapped on a curve because error detection thresholds are adjustable. For example, false positive rates can be reduced if system users are willing to accept higher false negative rates and vice versa. Probabilistic systems allow thresholds to be set to achieve the relationship of false positives to false negatives of any point on the curve.

All Probabilistic Systems are Not Created Equal

Not all probabilistic algorithms applied to the same set of circumstances yield results with the same degree of accuracy. Therefore, businesses should understand the features and capabilities necessary for accurate and effective automatic probabilistic data matching. The following are some of the most critical:

- **Dual thresholds**. Most systems based on probabilistic algorithms can be tuned to achieve specific false positive and false negative rates. However, not all provide the ability to set multiple thresholds for each search. This "dual threshold" capability is critical for businesses that need to ensure a very high degree of accuracy for certain matches, while a lesser degree for others. An example would be a state voter registration system where you would want a small false positive rate in looking for duplicate registrations but also require a low false negative rate when matching against felon lists.
- **Real-time response**. Not all probabilistic systems can produce results in real time. Businesses that require real-time capability should avoid solutions that offload batch processing with no emphasis on performance. Instead they should look for a system that can scale to support millions and billions of records for on-demand record lookups. To provide an up-to-the-moment view of customer data across the enterprise, as well as prevent duplicate records from entering the system on an ongoing basis, the ability to operate in real time is essential.
- **Adaptability**. Businesses concerned with high accuracy should also look for a highly adaptive system - one that adjusts according to the data contained in individual files. Such a system could adjust the matching algorithm as data quality changes in the underlying file.
- **Extensibility**. To ensure very high accuracy, companies must be able to include search parameters specific to their business or industry. Some probabilistic engines allow easy addition of new data fields without extensive business rule revision, thus

enabling businesses to add new parameters in accordance with changing search requirements. This would allow certain types of businesses and organizations (such as law enforcement agencies) that rely on a high degree of detail for identification to add more fields than the default number. For example, police departments could add fields to search by physical characteristic (i.e., tattoos or hair color).

Additional Considerations

Businesses in the process of evaluating probabilistic matching systems should assess whether the solutions meet all IT requirements:

- Utilize proven probabilistic algorithms that score and match data across a variety of attributes using likelihood statistical theory for the highest levels of accuracy.
- Do not require programming code to be added to source systems or modification or standardization of source data.
- Support core data from sources in original form and maintain complete historical versioning.
- Support multiple matching score thresholds to help manage quality versus cost tradeoffs.
- Include a complete task model to prioritize and resolve data errors or ambiguous linkages.
- Offer role-based security access down to the attribute level.

Picking the Right Solution

Although less accurate, deterministic matching can be an acceptable, lower-cost matching option for organizations that have smaller data sets, fewer attributes and require less complex rules. Others may favor probabilistic matching due to its flexibility and ability to check potential matches against a higher number of variables and very large and growing databases. Probabilistic solutions can provide enormous value for businesses in many industries, including healthcare, financial services, hospitality, and public safety, and for many applications, including customer relationship management and business intelligence. Businesses that require the highest levels of accuracy in real time, regardless of data volume, can fulfill their needs with a probabilistic solution, enabling them to limit the number of inaccurate matches and eliminate the costs and consequences associated with making them.

Scott Schumacher serves as chief scientist at Initiate Systems where he is responsible for R&D of Initiate's matching algorithms and the overall management of product development. Initiate Systems, Inc. is a leading provider of customer-centric master data management (MDM) solutions for companies and government agencies that want to create the most complete, real-time views of people, households and organizations from data dispersed across multiple application systems and databases.

BT&SG
Business Technology & Securities Group