INFOGLIDE SOFTWARE

White Paper

Methodology and Tools for Effective Data Quality Management

PN-10007-02

Table of Contents

Purpose

The purpose of this white paper is to discuss how Infoglide Software uses the Identity Resolution Engine (IRE) in order to resolve data quality.

Tool

Identity Resolution Engine™ (IRE)

The IRE is a unique enterprise solution that satisfies the growing need for identity resolution. IRE aggregates information from existing data stores to form a clear, comprehensive, composite depiction of an individual or entity. IRE glides across multiple data sources, applies sophisticated Similarity Search algorithms and techniques, and presents a unified view of an individual or entity. It is also able to highlight otherwise hidden relationships based on the similarity of multiple attributes.

IRE and Data Quality

IRE is designed to handle the following data quality issues:

- Multiple Data Sources – IRE performs identity analysis across multiple data sources such as white lists (allowed individuals), watch lists (disallowed individuals), historical data, and so on. The data may be located in one or more databases. Although IRE does not require data warehousing, it supports automated loading of data from extracts or other databases through its scripting facilities so data can be staged when appropriate.
- Disparate Systems – Searching across different systems of different types to return a unified result is a key strength of IRE. Information from heterogeneous Open Database Connectivity (ODBC) or Java Database Connectivity (JDBC)-compliant databases can be transparently "similarity searched" to resolve multiple identities and discover hidden relationships.
- Different Structures – Many times, the data contained in one database is in a different structure than in another (e.g., Full Name in one column versus First Name, Middle Name, and Last Name in three separate columns) or the names of the column headings are different (e.g., one table has the data for first names in a column called FName and another table has it in a column called First Name). IRE assists the matching process by mapping data with the same meaning across multiple databases on the fly. Infoglide Software's Similarity Search algorithms provide additional support by handling cases in which, for example, a middle name is missing, the first and last names are transposed, or initials are used. By making such associations, IRE can search data and find hidden relationships within multiple sources regardless of type or physical location.

- Dirty Data – Because of data entry errors or intentional data manipulations by criminals, finding matches within your data is often problematic or impossible with common search techniques. Many methodologies used to identify individuals in new data depend on finding exact matches within databases. This works fine in theory, but in reality, unless the data matches perfectly, these methodologies are ineffective and increase the rate of false negatives or incorrectly discarded records. In contrast, by using Similarity Search, IRE gathers up the most similar records, presents them for comparison and analysis, and identifies unintentional as well as intentional data entry errors in order to effectively stop sophisticated criminals from being able to fool the system.

Methodology

Figure 1 depicts the various layers that comprise IRE from the bottom layer of Similarity Matching all the way up to the top layer of Business Processing. Each of these layers is discussed in the sections below.

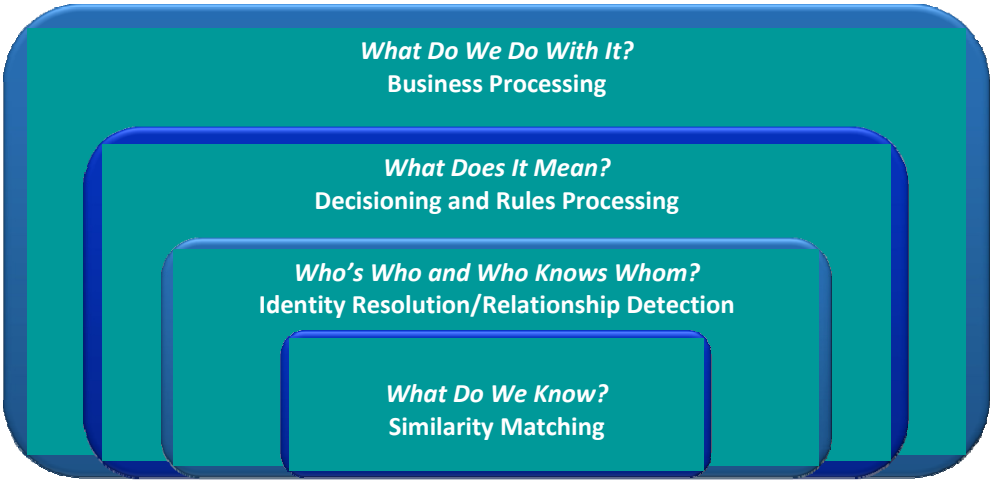

Figure 1: IRE Overview

Similarity Matching

By using Similarity Search, IRE gathers up the most similar records and presents them for comparison and analysis. IRE's patented similarity algorithms are designed to quickly and accurately return otherwise undetected results in real time. Matches can be made against cached data sets, live databases, local web services, or external data vendors.

Built into IRE are over 50 Similarity Search algorithms that calculate the distance between search and target attributes.

IRE's algorithms are capable of resolving the following:

- First Names
- Last Names
- Full Names
- Addresses (US, UK, French, and Irish)
- SSN
- DOB (American as well as International date formats)
- License Plates (domestic and international)
- Passport Numbers
- Telephones (US, Canada)
- Email Addresses
- URLs
- Phonetically similar words
- Lexically similar words
- Visually similar words
- VIN Numbers
- Credit Card Numbers
- DL Numbers
- Airport Names and Codes
- Country Names and Codes
- State Names and Codes (US, Canada)
- Time Representations
- Anagram Words
- Synonym Words (for colors, cars, drug names, or other categories)
- Telephones (UK)

Relationship Resolution

IRE's Similarity Search results are powerful and can answer the question, "Who's Who?" IRE can work in real-time or as part of an investigator's armory, but it is IRE's ability to detect non-obvious relationships and answer the question, "Who Knows Whom?" that helps differentiate it from similar solutions. It is imperative that effective identity resolution solutions uncover what a match between two records indicates – the same person, a shared household, or identity fraud – to know the appropriate actions to take.

IRE discovers hidden relationships between individuals by evaluating the degree of similarity between their attributes and extending the recognition of relationships beyond a single degree of separation. The combined activities of fraudsters who share addresses, phone numbers, and other personal data are readily exposed.

When domain-specific information is available, IRE is able to make even more sense of the matches found. For example, if matches with similar street addresses are determined to be of most interest then IRE can cluster matches with that type of relationship in order to effectively prevent fraud accomplished by manipulating or falsifying addresses.

IRE's web client includes a sophisticated relationship-charting capability (shown in Figure 2) that graphically displays hidden relationships among your data.

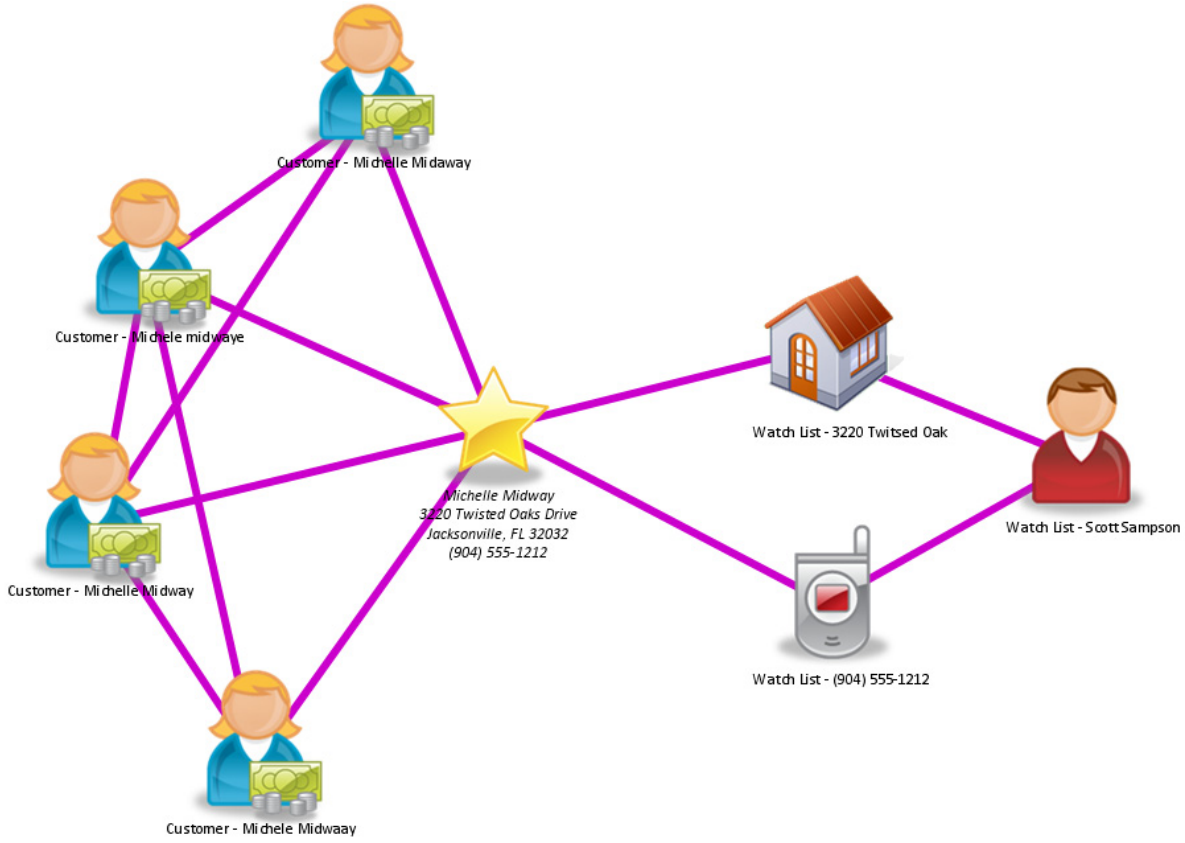

Figure 2: IRE Link Charting

The improved quality of relationship resolution intelligence provided by IRE simplifies business rules development and applications by reducing data "noise" and false positives in identity resolution results.

Decisioning and Rules

Building upon its identity intelligence, IRE offers a variety of configurable decisioning engines. Specific rules can be applied to the intelligence gathered from IRE's search, relationship, and identity results, enabling IRE to execute an explicit action based on the results of certain conditions.

For straightforward tasks, a scripting engine automates the evaluation of identities. In response to a detected condition, the engine can perform additional searches or decisions, add a record to a database, send a message via e-mail or message queue, or apply a label. For example, you can search a name and address against a list of known offenders and then have a custom rule that labels the results of the search according to how closely they match an entry in the list. In other words, an entry that strongly matches the offender list could be labeled as “red”.

More complex decisioning is supported by a pattern-matching rules engine with a rich, high-performance programming language environment. The engine makes complex reasoning – such as “flag any employee who shares a household with a customer who matches a known shoplifter” – straightforward to implement. A record of the decisions made is available from IRE’s audit log so that you can implement automated decisioning while also maintaining a traceable data trail.

All engines can be integrated using web service interfaces to enhance existing or new applications and business processes.

Business Processing

When provided with additional information and a better understanding of the business logic that needs to be applied, IRE can further classify whether a match is an attempt to deceive and apply appropriate business rules to perceived risks. In the example above where names that closely matched an offender’s list were labelled as “red” using IRE decisioning capability, IRE’s business processing could further define the rule so that all “red” requests are either tagged for later analysis, put on hold while the results are routed to a fraud team for further evaluation, or automatically denied service, depending on business requirements.

Process

Data Pattern Variations

The algorithms used to detect matches are able to detect numerous patterns in data. In order to provide a sample of the data patterns that IRE detects, we have highlighted the following data types:

- Name
- Date of Birth (DOB)
- SSN
- Street Address
- City
- State
- Zip

Note: The patterns provided in the sections that follow are only a sample of the numerous patterns IRE is capable of detecting. Since the algorithms used within IRE are purposely designed to be tuned to any data set IRE searches, the list of data patterns that IRE is able to detect is continually evolving and growing.

Name

Table 1: Sample Data Patterns for a Name Search

Data Pattern	Example(s)
Exact match	Jane Doe vs. Jane Doe
Case independence	John Smith vs. jOhN SmItH
Up to 2 inserted characters	• MCMAHON vs. MACMAHON • James Johanson vs. James Johanneson
Anglo nick name*	• Robert Witt vs. Bob Witt • Charles vs. Charlie
Missing Tokens	James Douglas Morrison vs. James Morrison
Arabic nick name*	Mohammad Atif vs. Muhamed Atif
Abbreviated names	• James D Morrison vs. J D Morrison • Company vs. CO • "and" vs. "&"
Switched characters	Jane Doe vs. Jaen Doe
Token switches	Osama Bin Laden vs. Bin Laden Osama
Pre-names	Claude Van Damme vs. Claude Vandamme
Character change	• Benny vs. Banny • Gregorey vs. Grigoriy • dental vs. deltam
Three character addition	FRAN vs. FRANCIS
Transpose first, last, and/or middle name	• Ho NHI vs. Nhi HO • Tom Allen vs. Allen Tom
Initials	V GHAZARIAN vs. Vanik GHAZARIAN
Punctuation (i.e., Hyphenated vs. Unhyphenated, acronyms with periods or without, apostrophes)	• Mary-am vs. Maryam • R.A.M. LLC DBA CATALINA MARKET vs. RAM LLC DBA CATALINA MARKET • Thanh N'GUYEN vs. Thanh NGUYEN
Spaces	ORIONGROUP INC vs. ORION GROUP INC
Repeated Word	Koon KOON SUH vs. Koon SUH
Numerals	Oracle, Inc. vs. 123 Inc.
Companies	• Oracle Inc vs. Oracle Corporation • Oracle vs. Oracle Corporation
Recognition of Titles	• Mr. Robert Johnson vs. Mrs. Robert Johnson **Note**: Examples of other recognized titles include: Mrs., Pvt., Dr., etc.

*Nicknames are checked using an externally pluggable cultural name matching synonym table. For more information on IRE's cultural name matching capability, see page 8.

Date of Birth (DOB)

Table 2: Sample Data Patterns for a DOB Search

Data Pattern	Example(s)
One number change	19510530 vs. 19510531
Two number change	19611107 vs. 19611110
Date format change	• 12/25/1999 vs. 12-25-1999 (different delimiters) • 12 December 1999 vs. 12 Dec 99 (short forms) • 12/25/1999 vs. 25.12.1999 (American vs. European)

Social Security Number (SSN)

Table 3: Sample Data Patterns for a SSN Search

Data Pattern	Example(s)
1 character change	123-45-6789 vs. 123-45-6785
2 character change	123-45-6789 vs. 144-45-6789
Character addition	123-45-6789 vs. A123-B45-C6789

Street Address

Table 4: Sample Data Patterns for a Street Address Search

Data Pattern	Example(s)
One number change	6318 vs. 6317
Two characters transposed in street name	Farman vs. Farnam
Character changes	• Breeman vs. Freeman • Cemetary vs. Seminary
One character addition	• PO Box vs. PO Blox • PO Box vs. PO Boxx
One character removal from street name	Hanloan vs. Hanlon
Abbreviations	• ln vs lane • S vs. Soth • Apt vs. Apartment
Comma added	1601 Elm Street, Suite 3000 vs 1601 Elm Street Suite 3000
Numbered Street Names	112 Second Street vs. 112 2nd Street

City

Table 5: Sample Data Patterns for a City Search

Data Pattern	Example(s)
One character change	Tavannah vs. Savannah
One character removal	Los Angels vs. Los Angeles
One character addition	ANAHEIME vs. ANAHEIM
Comma included	Vigo, Spain vs. Vigo Spain
Transpose tokens	Los Angeles vs. Angeles Los
Abbreviations	W. LaFayette vs. West LaFayette

State

Table 6: Sample Data Patterns for a State Search

Data Pattern	Example(s)
Two character removal	Texas vs. Tex
Abbreviations	Texas vs. TX

Zip

Table 7: Sample Data Patterns for a Zip Search

Data Pattern	Example(s)
Number switches	78721 vs. 78712
Number insertions	78721 vs. 787012
Number changes	78721 vs. 76721

Cultural Name Matching

Included in IRE is the capability to detect name matches within various cultures. Currently, IRE supports name matching for the following cultures:

- Arabic
- Czech
- Danish
- Dutch
- English
- Finnish
- French
- German
- Greek
- Hungarian
- Irish
- Italian
- Latin
- Norwegian
- Polish
- Portuguese
- Romanian
- Russian
- Scandinavian
- Scottish
- Spanish
- Swedish
- Swiss
- Welsh

Note: IRE supports name matching for other cultures in addition to those listed above, but those cultures that are more abstract were not listed. If a culture is not listed above, contact Infoglide Software to learn whether IRE supports that particular culture and how IRE would handle data from that culture.

In terms of data quality, cultural name matching accounts for discrepancies in data due to cultural variances and ensures that the data is scored accurately and not penalized for differences that are strictly cultural-based. For example, Table 8 illustrates the impact on a match's similarity score when IRE's cultural synonym table is applied versus when it is not.

Table 8: Sample Scores with Cultural Name Matching vs. Without

Culture	Name Matches	Score without Cultural Name Matching	Score with Cultural Name Matching
Dutch	Margriet Barhydt Gretchen Barhid	59%	86%
Italian	Giovanni Di Alberto Vanni Dialberto	75%	95%
Polish	Nikolaos Adamicz Mikolaj Adamics	76%	93%

As seen from the examples in Table 8, even though the source and target names may appear to be very different, IRE's cultural name matching capability is able to detect cultural similarities (i.e., Gretchen is a Dutch nickname for Margriet) that accurately raise a match's score.

Post Processing

Once IRE identifies a suitable candidate set using Similarity Search, a post-processing phase is implemented that handles data quality issues by performing an inter-attribute comparison and scores matches by using a comprehensive labeling and points system.

Inter-Attribute Comparison

IRE's post-processing phase can compare individual attributes between the source and target records. The inter-attribute comparison is fully configurable, which includes not only determining which algorithm to use but also which fields to compare. For example, given a first and last name on both records, IRE can be configured to not only compare First Name against First Name and Last Name against Last Name but additionally compare First Name against Last Name and Last Name against First Name to account for possible name swapping or transpositions.

Scoring

IRE also includes a labeling and points system that makes it possible to get a better idea of which results matched better than others and why records were classified as a match. Post processing, IRE classifies each match with a type code that identifies whether a record matched a particular attribute and to what degree (high or medium to high). Next, a point system assigns a numeric value to each type code. The point system is configured to give more points to those codes reflecting attribute matches. Finally, the composite points for each type code are grouped into ranges in order to indicate the overall confidence in a match.

Conclusion

In summary, unlike traditional technologies, IRE addresses data quality issues by providing:

- Support for multiple data sources, disparate systems, different data structures, and dirty data;
- Similarity Search technology that uses sophisticated algorithms to search across multiple attributes and data sources to provide a single view of an individual or entity;
- Relationship and identity resolution technology that automates, analyzes, and augments search results to uncover related individuals or entities;
- A Decisioning engine that uses customizable rules to apply business logic to both incoming data and search results in order to customize how an overall "decision" or "classification" is reached; and
- Identity intelligence that distinguishes matches and simplifies results using a post-processing scoring method.