IASIU SIU *eNews* 2Q2011

Featured Article

Entity and Relationship Resolution for Better Fraud Detection
By Nigel DeFreitas

Among the vast amounts of electronic data collected by insurance carriers are many duplicate records — claims data, policy data and a wide range of other information on individuals and companies that may or may not be the same entity. We rely on our computer systems to sort out who's who, then to accurately identify which individuals or businesses performed which claim transactions.

But what about if those computer systems aren't up to the task? What happens, for example, when a single policyholder repeatedly files multiple, questionable claims using the same attorney and chiropractor, but goes undetected by using different names, addresses and/or phone numbers? This system vulnerability opens the door to fraud.

A process known as entity resolution might be the best solution to the problem. Think of entities as the individuals or involved parties on a claim, or an organization such as a medical facility. Entity resolution is the process of comparing the data attributes of two or more individuals or businesses to see if they represent the same person or business. Sounds pretty simple, right? Not always. Humans can tell people apart fairly easily. However, finding matching data, then attributing it to a single individual or business can be a difficult task for computer systems.

Can computer systems behave like humans and recognize which businesses or persons are duplicates? If so, what are the implications for fraud analytics and decision management capabilities?

Gray Matters

In his book *On Intelligence* , Jeff Hawkins, creator of Palm computing, explains his mental model for how human intelligence and the brain work. Hawkins theorizes that the eyes, ears, nose and nervous system act as sensors that feed stimuli to the neocortex, which sifts through those stimuli to make sense of everything we see, smell, hear, taste, feel and *remember* .

The neocortex, according to Hawkins, is set up to filter those stimuli so we recognize significant events and match similar events from our memories with ease. It is how we instantly recognize a friend's face even after years of being apart and how we solve picture puzzles by looking for pieces with similar edge characteristics.

Humans compare facial traits and other characteristics of an individual to their memory of that person. Those characteristics might be as subtle as the cadence and sound of someone's footsteps. First, we hear the footsteps and think, "That sounds like Anna." Then, we peek out into the hallway and verify, "Yes, it is Anna." Hawkins refers to that cognitive process of prediction and verification as the *memory-prediction framework* . And it's something our brains do without having to scan through the list of everyone we've ever known until we reach the record matching the face in front of us.

The same can't be said for most insurers' computer systems, which must scan every record for matching entity attributes, such as first name, last name, Social Security number, email address, phone number and other information. Computers typically lack more definitive methods — such as biometrics (fingerprint scans, retina scans, or DNA sequence matching) — to zero in on a single record. Using fuzzy search methods, computers perform comparisons against many records in the database, then

return results to a human who makes the final determination. Scanning all records each time isn't scalable, and it's not how humans match people or puzzle pieces. We don't pay attention to every bit of stimuli our senses emit. We look for similarities, draw from our memories, then draw a conclusion.

It turns out , though, that similar processing strategies can be architected into our computer systems. Trouble is, current implementations often do a poor job of de-duping records into unique identities, so as new data arrives, the number of duplicates actually multiplies — the opposite the desired outcome.

Why Entity Resolution Matters

If our computer systems are unable to figure out who is who, then they won't be able to accurately identify which individuals or businesses performed which transactions. One example is the aforementioned policyholder who filed multiple, questionable claims using the same attorney and chiropractor, but who went undetected by using different names, addresses or phone numbers.

In July of 2010, the U.S. Department of Justice and Health & Human Services arrested 94 doctors, nurses and healthcare company owners for submitting $251 million in false claims. Those Medicare fraud schemes often involved setting up shell storefronts for short periods of time and submitting fake claims for expensive prosthetics using patient identities bought on the black market. It's common for fraudsters to submit multiple claims for the same patient involving three or four limbs.

Good entity resolution and alerts could have prevented this by recognizing and linking together seemingly distinct identities.

The Entity Resolution Solution

Similar to how look at puzzle piece edges or facial characteristics to determine if two pieces go together or if two people are siblings, computer systems can be designed to look at identity attribute information to determine the likelihood that two records represent the same individual or business. They can also be configured to explore these relationships, then remember them when similar identities are encountered.

A wide range of factors can be used in entity resolution, among them first and last name, address, email, Social Security number, Tax ID, date of birth and phone number. Resolution involving those factors can be pretty straightforward. For example, if two records share the same Social Security number and same last name but different first name ("James" and "J"), then they most likely represent the same person. Likewise, if the address and the phone number are the same but the name and social security number are very different, we could infer that the two identities are related. More information will yield a higher fidelity of resolved entities.

Unless a fraud perpetrator is using completely new identity attributes every time, it 's likely they will leave behind a trail that leads directly to their true identity. This applies to transposed digits, variations in names and other attributes. When this occurs, algorithms can be used to find similar but different attributes and offer close matches to help make the final determination.

Similar to the way humans categorize puzzle pieces or ignore most of the stimuli our senses emit, software systems can be designed to group similar entities and compare new records against the smaller subset of remembered identities. There's no need to take a boil-the-ocean approach to finding similar entities.

Fraud and the Fire Hose

Think of all incoming transactions as a fire hose of data that's pumping information into a safe harbor for fraudsters. These perpetrators are aware that if they use a significantly different name, address or other identity information, they likely won't be linked to identities on watch lists. They adapt, and they hide among the rest of the population of honest policyholders. Hence, the true cost of insurance fraud is frequently difficult to identify.

The quantity of claim and policy data the underwriters receive each day is often too great to be manually checked and verified for fraud. To compound the issue, as claims databases grow, so do the number of identities that we need to compare. Current models that rely on a "search everything" approach are not scalable, even for computer systems. It's more manageable to remember the smaller subset of similar identities or entities and update them with every new variation or alias by comparing only similar records as they arrive.

A lot of information arrives in the form of various daily insurance transactions, but identity attributes represent a smaller subset of that data. It's possible to leverage that stream of data using software design patterns (such as wire-tap) with staged event-driven architectures (SEDA) to triage the information. As a result, only the relevant identity and context data are passed to a central entity resolution system, similar to the way our brain filters out some of the vast stimuli our senses feed it. Now we're talking about "sipping from the fire hose" of enterprise data.

Potential issues

Entity resolution isn't without its shortcomings. Take, for example, cases involving naturally large entities such as multinational corporations that use the same tax ID and name. One side effect is that incoming entities with similar sets of characteristics will need to be compared to the large entity. This will require more computing resources, similar to when a human is faced with a larger set of very similar puzzle pieces. However, this is still less expensive than comparing the new entity to every entity on record.

Addressing false positives is another issue with automatic entity resolution. The process may sometimes entail a human decision before merging two entities. For that reason, it may require a significant amount of time and human resources — via follow-ups and fact-checking — to correct entities and relationships that have been incorrectly resolved. To avoid the need for manual intervention, the best approach may be to merge entities based on sound resolution rules and ignore close matches.

Though it's technically feasible to tap into streams of enterprise data, there may be contractual obligations regarding how that data can be used. Some contracts may prohibit the comingling of data, for example, while others may require data to first be anonymized before it is repurposed. To ease the process of adding new data sources to your entity resolution system, employ an enterprise data management group.

Conclusion

Traditional de-duplication techniques are inferior to holistic entity and relationship resolution and lead to improper accounting of identities in most computer systems. As a result, many insurers may release payments on fraudulent claims. Employing proper identity resolution and data governance programs will enable better fraud prevention and more scalable solutions as data grows. Humans recognize similar entities with ease. Shouldn't our computer systems, too?

Nigel DeFreitas is the chief application architect at ISO in Jersey City, N.J. He is currently investigating entity resolution and predictive model deployment systems — technologies that help identify fraud and boost decision-management capabilities. ISO is a leading source of information about property/casualty insurance risk.

Back to SIU eNews