The Challenges of, and Why You Should Reconsider Using, Truth Sets for Optimizing Entity Resolution: A Case Study

(Completed Research Paper)

Pei Wang, PhD[1], Daniel L. Pullen, PhD[2], Maryam Y. Garza, MMCi[1] and Meredith N. Zozus, PhD[1]
[1]University of Arkansas for Medical Sciences College of Medicine;
[2]Black Oak Analytics
pwang@uams.edu, dpullen@blackoakgroup.com, mygarza@uams.edu, mnzozus@uams.edu

Abstract: The creation of high quality Entity Resolution (ER) processes depends on the ability to quickly and effectively identify erroneous outcomes (false positives and false negatives) in ER results. In past and current research, truth sets have been used to provide this ability. Unfortunately, managing the quantity of data provided to reviewers for manual annotation during the generation process often forces researchers to generate sampled data that is not entirely representative of the total amount of variation contained within the original dataset. This often causes an over-fitting of the match logic to the truth set. This case study shows the challenges and issues that can arise when using truth sets for creating and analyzing ER matching logic.

Keywords: Truth Set, Entity Resolution, Boolean Match Rule, EHR Data

BACKGROUND

Entity Resolution

Entity Resolution (ER) is the process of determining whether two references to real world objects in an information system are referring to the same object or to different objects (Talburt 2011). The references are made up of attributes and the values of the attributes describe the real-world entity to which they refer. The ER processes discussed in this paper uses Boolean match rules to make decisions. Boolean match rules do not produce a score or weight when comparing a pair of references, only a True/False decision. If two references satisfy a Boolean match rule, the references are linked together. After the application of transitive closure, all of the references that can be linked together have been matched and form an entity identity structure (EIS) (Zhou et al. 2011). An EIS is often labeled as a cluster in ER literature.

Boolean Match Rules

Boolean match rules are used to determine the outcome as "link" pairs or "non-link" pairs. The basic unit of a Boolean rule is a term. In mathematics, the term is typically referred to as the predicate. A term is the comparison between the values of an attribute in the pair of records. For example, a term could compare first names or dates-of-birth. The term is considered to be “TRUE” if the degree of similarity required by the comparison is met. The similarity measure is given by a similarity function such as the Levenshtein edit distance or Soundex code. The rule itself is made up of a series of terms connected by “AND” logic, i.e., every term must be true in order for the rule to be true. Likewise, Boolean rules can also be connected by “OR” logic, i.e., the pair of references should be linked if at least one of the Boolean rules is true (Fellegi et al. 1969).

In evaluating the outcome of an ER process, the results of the matches between all pairs of references can be categorized into four outcomes: true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). TPs are correctly labeled "link" pairs. TNs are correctly labeled “non-link” pairs. In contrast to these correct results, there are two types of incorrect linking results. FPs are pairs of records that have been identified as matches or “link” pairs by the ER process but actually refer to two different real-world entities. FNs are pairs of records that have been identified as non-matches or “non-link” pairs

by an ER process but actually refer to the same real-world entity (Christen 2013). The goal of an ER process is to produce the lowest number of FPs and FNs.

The OYSTER ER System

The ER processes in this paper were performed with OYSTER (Open sYSTem for Entity Resolution). OYSTER is an open source ER system developed by the Center for Advance Research in Entity Resolution and Information Quality (ERIQ) at the University of Arkansas at Little Rock (UALR). It was specifically designed to support Entity Identity Information Management (EIIM) (Zhou et al. 2011). Although OYSTER can be run in several different configurations to support the various phases of the entity identity information life cycle, only the identity capture configuration was used for the results given in this paper (Zhou et al. 2012). To perform the ER, the OYSTER system was configured with 1) the data elements used to match the records and 2) the algorithm(s) by which to match, and then tested to identify the combination that yielded the optimum balance between proportion of FP and FN.

Truth Sets

Truth Sets are a collection of records, typically sampled from a larger data source that are systematically generated and analyzed by a human. They provide properly annotated linkage for their records. Although there are methods of estimating the correctness of a matching process, comparing the results against matches known to be TPs or TNs in the real world is the only truly definitive way to measure the accuracy of a matching process. Existing estimation methods use statistical processes for sampling portions of the data to generate an estimation of the errors and non-errors of a matching process. More sophisticated contemporary methods use clerical review indicators in combination with stratified sampling and estimation processes rather than pure random sampling (Pullen et al. 2011). These methods are used in a formative manner to optimize accuracy of matching, and in a summative manner to report match accuracy along with match results. Unfortunately, for much research based on record linkage, little summative detail is provided about the accuracy of the matching (Jurczyk et al. 2008). Furthermore, datasets with known matches and non-matches, i.e., truth sets, are rarely available for use in a formative or training manner (Christen 2008; Washio et al. 2008). This requires that researchers generate truth sets on demand for each dataset used.

CASE DESCRIPTION

The research setting

The MURDOCK Registry is a longitudinal, community-based health study collecting data from consenting adult residents from Kannapolis and Cabarrus counties (and surrounding regions) in North Carolina [12]. All participants enrolled in the study provide self-reported data (updated annually) and consent for longitudinal access to their electronic health records (EHRs). The objectives of the study required that the independently collected self-reported dataset be combined with (or linked to) the EHR data obtained from multiple facilities to characterize the accuracy of the EHR versus the self-reported data. In order to perform the characterization, participants would be contacted and interviewed by phone to discuss the identified discrepancies from the various data sources. In order to minimize the errors in participant identification, the accuracy of the data linkage needed to be assessed. Therefore, a truth set was created for optimizing the matching process according to match accuracy, so as to minimize FPs and concretely report the accuracy of the matches.

The truth set

It follows what is known about ER, that a truth set should be as representative as possible of the data on which the optimized rules will be run. The challenge lies in designing a truth set that is as similar as possible to the actual dataset because the linkage is dependent on the type of data available. If not

carefully considered, the FP and FN rates derived from the Boolean rules designed for the truth set may not necessarily translate well to the full dataset. For this reason, a truth set was generated using a subset of participants from the self-reported dataset and an EHR dataset from a small, local, outpatient clinic.

At the time of truth set creation, the registry database contained 10,069 participants, all of which had consented to linking their EHR data to the self-reported data within the registry. The community clinic EHR dataset contained 25,924 reported individuals. (Later through the record linkage, several instances of split charts – data for the same patient stored under two separate charts – were identified). To create the truth set, a low-sensitivity match using a 50% confidence setting was performed between the two data sources using DataFlux (SAS, Cary NC). The lowest confidence was used in order to identify all potential matches with a remote likelihood of being a TP. At that time, a little over 200 of the participants had reported the community clinic as their primary care facility. The loose (50% confidence) match generated 1,621 records consisting of records from the self-reported data and the EHR data. There were 680 clusters in the truth set. From data profiling of the truth set (Table 1), it is apparent that attributes First Name, Last Name, DOB, Physical Street Address, Physical City, Physical State, and Sex have low percentages of blank values and high completeness. Thus, these attributes are better suited for use in linking the data.

	null Count	Distinct Count	Unique Count	Blank Count	Total Records
last name	0	299	25	0	680
first name	0	269	13	0	680
middle name	0	23	0	313	680
suffix	0	3	1	674	680
race	0	8	2	344	680
sex	0	2	0	0	680
data of birth	0	346	22	0	680
physical street address	0	501	341	0	680
physical secondary street address	0	24	19	652	680
physical city	0	25	13	1	680
physical state	0	6	5	1	680
physical postal code	0	29	13	340	680
mailing street	0	327	315	340	680
mailing secondary street	0	19	17	661	680
mailing city	0	23	13	340	680
mailing state	0	6	3	340	680
mailing postal code	0	28	14	340	680
home phone	0	218	205	451	680
work phone	0	49	47	631	680
mobile phone	0	227	223	451	680

Table 1: Data Profiling of the Truth Set

These initial matches (680 clusters) were then reviewed by local study coordinators familiar with the participants. A total of 340 definite matches were identified based on study coordinator knowledge of the participants. All other initial matches were confirmed or refuted by contacting the study participant who originally self-reported the data. Questions such as, “Have you ever had a visit at the clinic?” “Do you remember visiting the clinic last February?” were used for confirmation or refutation of matches. A total

of 36 participants could not be contacted. This resulted in those clusters being classified as unconfirmed. From the remaining 644 clusters, a set of 340 confirmed matched clusters and 304 clusters confirmed as non-matches were properly annotated and provided for use in optimization of the match rules.

BOOLEAN RULE DESIGN

Based on the data profiling of the truth set, multiple rules and probabilistic approaches were developed and tested. Six rules were chosen based on optimization between low FP and FN errors with the priority weighted toward lowering FP matches (Table 2). Three similarity functions are used in the chosen rules, Scan, Soundex and Normalized Levenshtein Edit Distance (LED).

Scan is a multipurpose similarity function (Pullen et al. 2013). It performs transformations on the input strings based on the parameters passed to the function. Scan can be used to overcome a variety of data quality problems. It includes the capability to filter all special characters and only include letters or alphanumerical characters. The scan function can reorder strings, read them from right to left or left to right, and perform transformations regarding the casing of alphabetical characters. In this rule set, the scan function was used on the attributes *First Name, Last Name, City, Address* and *Sex* to scan from left to right, keep alphanumeric data types, and change all alphabetical characters to upper case. The attribute *Date of Birth* was adjusted to keep only the numbers in the string (and remove any hyphens or slashes).

Soundex is a phonetic function used to associate two strings with similar pronunciation together. For example, consider that value1 ="Damieva" and value2 = "Dameiva." These two values will produce the same Soundex hash value, creating a match or positive outcome.

LED is a distance-based function that can help solve typographical errors by calculating the similarities between two words. For example, consider that value1 = "Mariah" and value2 = "Miriah". If the LED threshold is set to 0.83, the two values will be matched together.

	First Name	Last Name	Date Of Birth	Address	City	Sex
Rule1	Scan	Scan	Scan(Keep first 5 digit)			
Rule2	Scan	LED(0.5)	Scan			
Rule3	Scan	Scan		Scan		
Rule4		Scan	Scan		Scan	Scan
Rule5	Soundex		Scan	Scan	Scan	
Rule 6	Scan			Scan(Keep first 4 Alpha)	Scan	

Table 2: Boolean Rule Set

After applying the six rules to the truth set, the Talburt-Wang Index (TWi) was used to measure the similarity of two partitions (Talburt 2011). The precision, recall and F-measure was used to measure the accuracy of the results generated by the rules. Precision is the fraction of relevant instances among the retrieved instances, and recall is the fraction of relevant instances that have been retrieved over total relevant instances. F-measure is the harmonic mean of precision and recall (Precision and recall 2017). Through the comparison of the ER result against the truth set, three false negative clusters and zero false positives were identified. The best TWi between the truth and the results of the Boolean rule was 0.996. The best precision was 1.000, the recall was 0.991 and the F-measure was 0.995. Using the full EHR dataset added a maximum of 11 FP errors. This foreshadowed that adding more data would degrade performance, but still within acceptable limits.

After applying the six rules to the full EHR and self-reported datasets, clerical review indicators were run. Clerical review indicators are signals generated from the system, or a post-processing tool, that notifies users of potential FP errors and potential FN errors (Talburt et al. 2015). The initial run of the review indicators produced over one thousand clusters to be reviewed. Upon manual inspection and consultation with the local study coordinators, FP and FN errors were identified at higher frequencies than were indicated by initial testing and tuning with the truth set.

Upon comparison of the records and clusters, it was confirmed that the results seen in the review indicators were legitimate FP and FN errors. After much deliberation by the project team, over-fitting was identified as the root cause. By starting the development of the truth set with a loose match (50% confidence in Dataflux), the records had been narrowed such that the rules generated based on the truth set had not performed as well on the larger volume of data compared to our overly constrained truth set. This is a core limitation of using truth sets for matching logic development and a concern at the outset of this project.

DISCUSSION

The loose match provided a high number of true matches, almost fifty-fifty, with which to test the rules, while constraining the number of clusters that had to be manually identified. This significantly decreased the workload required for identifying matches. (Contacting the participants to confirm clusters took two study coordinators an elapsed time of three months at 20%-50% effort.) The high proportion of matches from the initial matching to create the truth set was helpful in giving us enough true match clusters to develop and test the rules. The constraint of the initial loose match also helped reduce the staff hours needed to develop the truth set. However, narrowing the records used for rule development in this way resulted in over-fitting the rules, i.e., tailoring rules to an artificially constrained set of input data such that the resulting rules did not account for the variability seen in the full dataset. Thus, the rules did not perform well over larger and more varied datasets.

The issues that were identified in the undertaking of this project demonstrate to the limitations of using a loose match as a starting point for truth set development. Any constraint of the features or variability in the actual data adds risk of developing inadequate rules. Thus, data profiling beyond null, unique and distinct is likely needed and should include measures of central tendency and dispersion for each data element evaluated as a potential match field. Furthermore, it is likely that additional measures of the variability in the data may not be enough to mitigate these issues. Unfortunately, this presents a zero-sum game—a proverbial "yin and yang" of truth sets. Increasing the recall mechanism of the truth set generation process to mitigate the over-fitting issue would dramatically increase the burden of review on the study coordinators to a point that the manual review process would consume an unreasonable or unattainable amount of time and resources. In contrast, using the original configuration of the recall mechanism of the truth set generation process produces an over-fitting of the matching logic to the truth set.

After this exercise, it is recommended to use independent algorithms such as those in clerical review indicators, as these provided an evaluation of the record linkage results with enough independence to identify clusters that were likely problematic. It was analysis of these clusters that led to the detection of the problem. Furthermore, comprehensive use of clerical review indicators with effective stratified sampling approaches can provide a mechanism for estimating the quality of much larger datasets with a high degree of accuracy.

CONCLUSION

In conclusion, it is reiterated that caution should be taken against the constraint in truth set development and recommended that, if possible, the truth set is a full-featured subset of the actual data for which ER is

to be undertaken. It is important to note that creating a truth set can be resource intensive. Balancing the need for enough true and false matches with full-feature coverage will likely increase the resources needed to create truth sets. This limitation has the potential to make the effective generation of a truth set unfeasible in practice. The ever-growing size of contemporary datasets and the variety of data sources available to organizations only further exacerbate these challenges.

ACKNOWLEDGEMENT

The research described in this paper has been supported in part through Contract Number: ME-1409-22573 from the Patient-Centered Outcomes Research Institute (PCORI), and institutional commitment from Duke University to National Library of Medicine grant 4R00-LM011128.

REFERENCES

Christen, P. 2008. "Automatic training example selection for scalable unsupervised record linkage", in Advances in Knowledge Discovery and Data Mining, pp. 511-518.

Christen, P. 2014. Data Matching, (1st ed.) Berlin: Springer Berlin.

Fellegi, I., and Sunter, A. 1969. "A Theory for Record Linkage", Journal of the American Statistical Association (64:328), p. 1183(doi: 10.2307/2286061).

Jurczyk, P., James, J., Lu, L., Janet D, C., and Adolfo, C. 2008. "FRIL: A tool for comparative record linkage", in AMIA.

"Precision and recall." 2017. Wikipedia, Wikimedia Foundation, June 26 (available at https://en.wikipedia.org/wiki/Precision_and_recall#F-measure; retrieved July 24, 2017).

Pullen, D., Wang, P., Talburt, J., and Wu, N. 2013. "A False Positive Review Indicator for Entity Resolution Systems Using Boolean Rules", in The 18th International Conference on Information Quality, .

Pullen, D., Wang, P., Talburt, J., and Wu, N. 2013. "Mitigating data quality impairment on entity resolution errors in student enrollment data," in Proceedings of the International Conference on Information and Knowledge Engineering (IKE). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Talburt, J. 2011. Entity resolution and information quality, (1st ed.) San Francisco, Calif.: Morgan Kaufmann/Elsevier.

Talburt, J. and Yinle, Z. 2015. Entity Information Life Cycle for Big Data, San Francisco, CA: Morgan Kaufmann/Elsevier.

Washio, T., Inokuchi, A., Suzuki, E., and Ting, K. 2008. Advances in Knowledge Discovery and Data Mining, (1st ed.) Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Zhou, Y., and Talburt, J. 2011. "Entity Identity Information Management (EIIM)", in International Conference on Information Quality, , pp. 327-341.

Zhou, Y., Talburt, J., Kobayashi, F and Eric, D. N. 2012. "Implementing Boolean Matching Rules in an Entity Resolution System using XML Scripts". Information and Knowledge Engineering.