Inferred Error Rates for Entity Resolution in Healthcare

(Research Paper)

Melody Penning, PhD
University of Arkansas for Medical Sciences
MLPenning@uams.edu

Abstract: Today we have large numbers of data sources in healthcare that have useful information, which in combinations can provide many times more information value than the sources individually. Combinations are based on mappings of one or more elements from each source to the other called entity resolution (ER) or record linkage. Doing ER correctly requires understanding the sources, carefully planning the resolution procedure, and testing the results. The problem with testing is knowing the results are correct. The following study describes an innovative approach to estimating error rates for ER using techniques already in use in information retrieval (IR).

Keywords: Entity Resolution, Record Linkage, Healthcare

INTRODUCTION

Today we are fortunate to have large numbers of data sources that have useful information, but unfortunately they are limited in scope, often to a particular business area. Combining these data sources can provide many times more information value than any individual source. The combination is based on a mapping of one or more elements from each source called entity resolution (ER) or record linkage. Doing ER correctly requires understanding the sources, carefully planning the resolution process, and testing the results. The problem is knowing when it's done correctly. The following study describes an innovative approach to estimating error rates for ER using techniques already in use in information retrieval (IR). Entity resolution is a need that arises in all areas where information is stored from marketing to electronic health records (EHR). The first academic reference to record linkage is from Robert Dunn M.D. who argued the value of constructing health information vital statistics for the United States.

The current methods to address the problem are truth sets, benchmarks and surrogate truth sets. Building Partial Truth Sets or Gold Standard Records are known true links. These are however only a subset of the population of interest and may not contain some error types. Benchmarks are the result of comparing to the results of a previous ER system that was considered trustworthy. However, matching differences are not necessarily errors. Finally, pseudo-truth or surrogate truth are areas of high confidence designated as pseudo-true match and non-match (Winkler 2007). This type of ER assessment is based on the idea that "Given two entity references, the more similar the values for their corresponding identity attributes, the more likely the references will be equivalent"(JRTalburt 2015). These are prone to the same weaknesses as benchmark sets though.

So the problem is how to find the error rate without a verified truth set? To solve this I have borrowed techniques from the information retrieval community since they face the same challenges, and solved them by using "inferred" measures which have been well tested (Voorhees & Hersh, 2012).

BACKGROUND

Linking records has been a topic of research since at least the mid-1940s. In 1946 Halburt Dunn M.D., the chief of the US National Office of Vital Statistics, wrote about building a 'book of life' for each citizen that should "start with birth and end with death" (Dunn, 1946). His reasoning was that the accuracy and completeness of vital records could be improved with this indexing. He admits he was inspired by the Canadians who already had a process like this in place. Dunn doesn't mention inspecting linkage results but in 1959 Howard Newcombe, a Canadian geneticist, picked up this thread and specifically discusses the reliability of the linkages. He worked to identify both false positives and false negatives. These efforts were now feasible due to the advantage of punch card data that could be fed into a computer for a sort before a visual inspection (Newcombe, Kennedy, Axford, & James, 1959).

This line of thought is adopted and improved by Fellegi and Sunter in 1969. They developed a formalized mathematical model of record linkage that used the strength of match results to limit the number of records which would need an inspection. The expected outcome was a limited occurrences of false positive and false negative errors (Felleg & Sunter, 1969). Following this work, ER evaluation measures have continued to use the values of false positives and false negatives common in predictive analytics.

Current research, however, indicates potential weaknesses with this methodology. John Talburt, mathematician and information quality (IQ) expert, points-out that this approach to accuracy cannot be effective if the error rates are unknown but that few data professionals endeavor to take systematic measurements. The difficulty is that in order to count the errors, you need to know which matches are true and which are spurious (Talburt & Zhou, 2015). Many researchers are attempting to address this problem using truth sets and/or benchmarking. A truth set is simply a dataset that has been verified to be correctly matched. In most cases this is not available so methods of truth set generation have been suggested involving either manual review of real data or creation of synthetic data with predetermined match status (Christen, 2012). William Winkler, principal researcher at the U.S. Bureau of the Census, observes that labeled data of this type is nearly impossible to determine in real-world situations. He attempts to get around this problem by using 'pseudo-truth' data as a stand-in for an actual truth set (Winkler). Benchmarking, on the other hand, is a comparison of current ER results with previous ER results for the same data set (Talburt & Zhou, 2015) (Christen, 2012). Neither of these approaches deliver consistent and dependable results. Verified truth sets will be small due to the labor involved in gathering them and benchmarked sets lack precision, because previous errors will just be carried forward (Talburt & Zhou, 2015).

METHODS

Current ER Approaches to Quality Assessment

ACCURACY

Fellegi and Sunter used classification to decide on records linkages, but clustering is another technique that is available. In 1971 William M. Rand's exploration of clustering evaluation resulted in the Rand index. The Rand index is used as an accuracy measure and provides a good general comparison of the portion of correctly assigned results to those that were incorrectly assigned (Rand, 1971). Where true positives are TP, false positives are FP, true negatives are TN and false negatives are FN.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

F-MEASURE

"A single measure that trades off precision versus F MEASURE recall is the F measure, which is the weighted harmonic mean of precision and recall" (Manning, Raghavan, & Schütze, 2009). Precision tells us how many retrieved are actually relevant TP/(TP+FP), and recall tells us how many relevant were returned TP/(TP+FN).

$$FMeasure = \frac{2(Precision * Recall)}{Precision + Recall}$$

TALBURT WANG INDEX

The measures mentioned above are currently used in ER along with the Talburt-Wang Index which provides a method of ER system comparison that is more efficient since is does not require the TP, FP, TN, and FN counts. Instead only the counts of the partitions, or link groups, and their overlaps are needed.

$$TWi(A,B) = \frac{\sqrt{|A| * |B|}}{|V|}$$

Where A and B are the counts of linked groups resulting from two ER systems, and V is the number of overlaps between the systems. This results in a possible range of values from 0-1, with 1 for only exact match of ER results (Talburt & Zhou, 2015).

Current IR Approaches to Quality Assessment

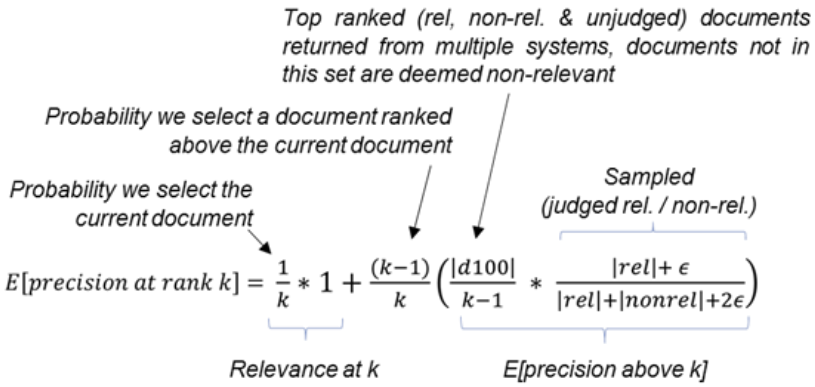
MEAN AVERAGE PRECISION

Mean average precision (MAP) is the mean of these average precision values and is helpful when a single number is needed to compare systems (Manning, Raghavan, & Schütze, 2009). The MAP score is about equal to the area under the precision recall curve. This measure is consistent across multiple systems using a common set of queries. "Among evaluation measures, MAP has been shown to have especially good discrimination and stability." (Manning, Raghavan, & Schütze, 2009).

The IR community has recently pushed evaluation into inferred measures to address the challenges caused by determining relevance within very large data sets (Yilmaz & Aslam, 2006) (Yilmaz, Kanoulas, & Aslam, 2008) (Voorhees & Hersh, 2012).

INFERRED AVERAGE PRECISION

An inferred version of AP was developed in 2005 and quickly adopted by the Text Retrieval Conference (TREC) run by the U.S. National Institute of Standards and Technology which has been a test bed for evaluating IR systems since 1992 and is well respected in the IR community. Here Average Precision is seen as the expected outcome of a random experiment where you select a set of relevant documents at random and let their ranks be k. For each sampled value, this equation will be repeated to calculate expected precision values at those ranks. The precision values can then be averaged just as in the descriptive version of AP to produce an inferred MAP value.



REL, NON-REL

The returned set contains documents which will be correctly returned (rel) or incorrectly returned (non-rel). For ER these are the matches and non-matches respectively.

THE d100 POOL

To eliminate the need to go through EVERY document to determine relevance the pool to be judged is limited and is typically about 2/3 of its possible size. In the ER case the d100 will always be k-1 since we cannot return 'non-d100' therefore all ER cluster pairs will be part of the d100

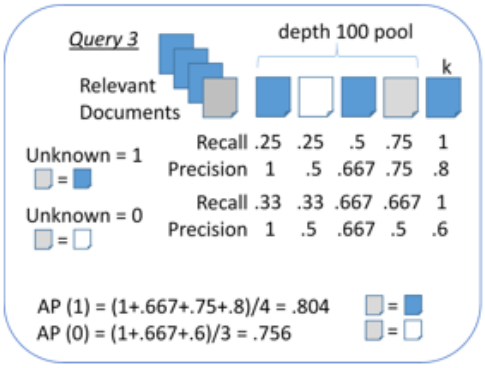
PRECISION @ K

The k value can take on two different types of values:

1. k can be at the actual cutoff which is represented in the first part of the equation. If so the relevance must be 1 since we chose k from only relevant documents.
2. k could also be ranked above the cutoff represented in the second part of the equation. The relevance here is determined by the probability of being in the d100 pool combined with the probability of being relevant at all.

Lidstone smoothing is applied in the small value added to rel and non-rel to prevent the possibility of 0/0 condition arising from no relevant or non-relevant documents above k.

In this example relevant documents are those in positions 1, 3 and 5. The document in position 2 is non-relevant and the document in position 4 is of unknown relevance. Precision and recall are calculated at each ranked position and AP is calculated twice, once assuming the unknown document is relevant and once assuming it is non-relevant resulting in AP values of .804 and .756. These values demonstrate how the infAP value of .756 is used as an estimator for AP. Because this example is so small sampling is done at every relevant k, for a large set sampling is done as needed.



$$E[precision\ at\ rank\ k] = \frac{1}{k} * 1 + \frac{(k-1)}{k}\left(\frac{|d100|}{k-1} * \frac{|rel|+\epsilon}{|rel|+|nonrel|+2\epsilon}\right)$$

E[precision at rank 5] = (1/5 * 1) + ((5-1)/5) * (4/(5-1) * (2/4)) = .6
E[precision at rank 3] = (1/3 * 1) + ((3-1)/3) * (2/(3-1) *(1/2)) = .667
E[precision at rank 1] = (1/1 * 1) = 1

infAP = (1+.667+.6)/3 = .756

Adapting IR Measures for ER

Adapting the IR measures for ER requires addressing ranking, incomplete truth sets and low error rates. In order to make this transition a query must be viewed in terms of a cluster of ER results. In making a query to any search engine you are requesting information, often documents, to be returned if they are related to the query test. A parallel to an ER cluster can be drawn here since an ER cluster is information about a single entity, often a person. A complete ER set would represent a grouping of all documents in a corpus according to some preset information needs or queries. This comparison can be further supported by the fact that when we use the results of ER we typically call for a cluster about one entity or group of similar entities to determine something about those real world entities.

Ranking can be achieved through a pairwise comparison of the members of a cluster. In order to quantify the ER cohesiveness, the cluster members can be scored on the goodness of the match. Ranking can be addressed by using pairwise match strength to score each pair. The scores can then be averaged for each element and the elements ranked according.

There are many possible approaches to scoring. One straight forward method is assigning points to be summed up to a total score for each equivalent attribute. Equivalence can be determined by matching algorithms just as in ER where exact match could be given a higher value than a partial string match. This is the approach taken in this research since the interpreting the scoring results is straightforward. Since this is a similar process to ER some overlap of rules should be expected. This design could present a problem if the scoring/ranking rules were exactly the same as the ER rules in the case of rule based ER. In this, perfect overlap situation the ranking of the cluster members would be a tie and ranking could not be completed. To avoid running into a problem of perfect overlap the scoring rules should be designed to use as many attributes as possible with high enough matching requirements to provide a good diversity of scores. In this research exact matches were scored at 1 and approximate matches were scored at .5 for all data elements. All measures were computed on same datasets.

TESTING DATA AND PROCESSING

The ER data used in this project is synthetic, created by software not as a result of real world data capture, but was generated using real identities. It was originally more than 20,000 true clusters made up from 271,143 rows. In order to test different aspects of the performance of an inferred measure against the performance of count based measures test runs were done using a subset of the synthetic ER data, the true link set and eight flawed links sets. The flawed sets were gathered using multiple runs of the OYSTER entity resolution system with different combinations of rules: first name only, first and middle name, address city and state, last name and state, last name and SSN, first and last name and address with last name SSN, first and last name with first name and phone, and finally first and last name with last name address city and state. Each flawed set was tested against the truth set and the measure results were recorded into result sets.

The intersection of the truth set links with the flawed ER links is composed of congruent and intersecting sets. Congruent sets are those which both ER systems agree on and intersecting sets are those which have incomplete agreement. To consider this factor, sampling for the inferred measures was tested both randomly and proportionally. The proportional samples were pulled from the congruent and intersecting sets in proportion with their occurrence. Both the random sample set and the proportional sample set were drawn and tested for each of the eight flawed ER test sets.

The results data sets built during testing consisted of four tables. Two tables were built at the cluster level (cluster level precision and inferred average precision), one table was built at the multiple cluster level and one table was built for error tracking.

SAMPLING

Stratified sampling uses some characteristic of the population to separate it into groups and then those groups are sampled from proportionally, whereas random sampling simply pulls samples at random from the population. Random and stratified, or proportional, sampling were tested on the same sets in order to experimentally verify the feasibility of different methodologies and to demonstrate the effects on the results. The proportional samples were selected based on the known agreement with the truth set. The clusters where the ER set agreed with the truth set were one group and the clusters where the ER set disagreed with the truth set were a second group.

ER clusters were used for sampling and testing instead of individual rows since in the ER context we are measuring the quality of the grouping not a characteristic of the data itself. Sampling was done of ER clusters in counts of 25, 50, 75, 100, 125, 150, 175 and 200. There were a total of 200 clusters in the test set. Each cluster count was sampled 30 times for each variation in k and cutoff. This resulted in test sets of 1920 per test set; 30 sampling runs x 2 cutoff x 4 k x 8 sample sizes = 1920 measure results for each ER set. For a real world situation Winkler recommends a minimum sample size of .5% of the sampled pairs to be tested (Winkler, 2007). For this ER set this would be close to 50 clusters since the average cluster size from empirical testing is ≈ 11.

Table 1

Clusters	Ave Pairs Per Cluster	All Pairs	.5% of Pairs	Number of Clusters
200	55	200 x 55 = 11000	.05 x 11000 = 550	550/11 = 50

RESULTS

Sampling proportional vs random

Because ER sets can be so large and because the errors can be limited to relatively few clusters it can be challenging to get a good picture of the error rate for the ER set using sampling. The chances of sampling an erroneous cluster is small in cases like these. To understand the effects of this issue proportional sampling is tested against random sampling.

To quantify the difference between random and proportional the variances at different sample sizes for the same ER set can be compared. If convergence is more rapid for proportionally sampled clusters then further testing of this type of sampling could prove valuable. The variance decreased equally from small to large samples for both randomly sampled and proportionally sampled clusters as shown in figures 1 and 2. This decrease in variance and associated interquartile range (IQR) is expected for larger sample sizes since IQR is a measure of variability and the value of the variable converges to the actual value as the sample size increases.

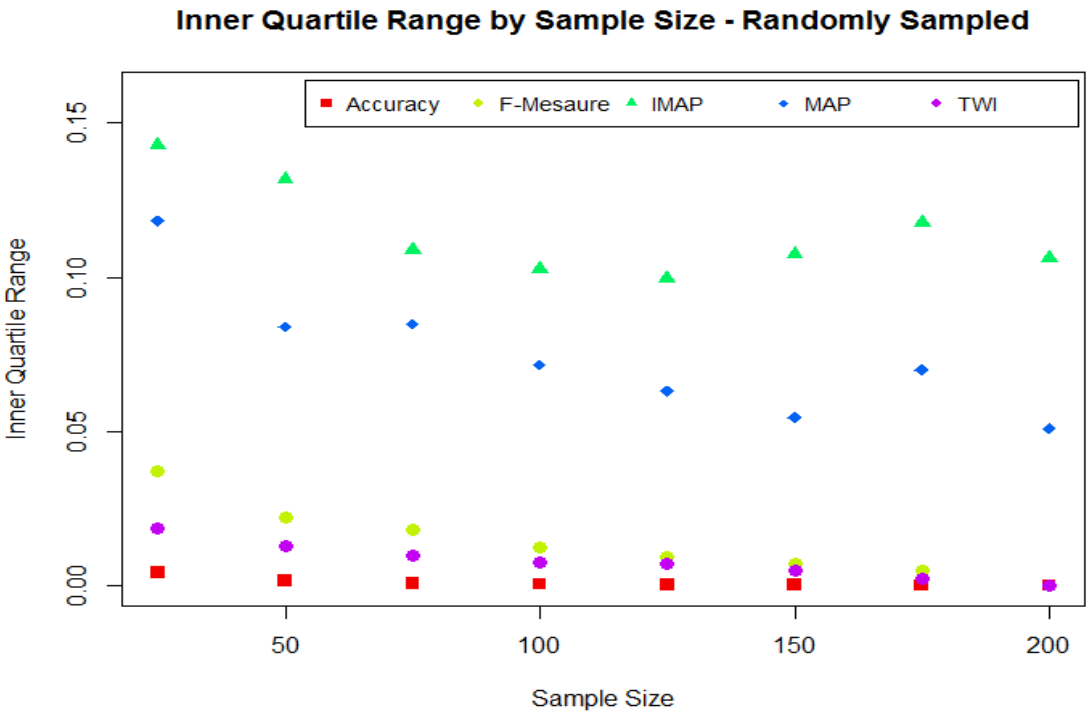


Figure 1

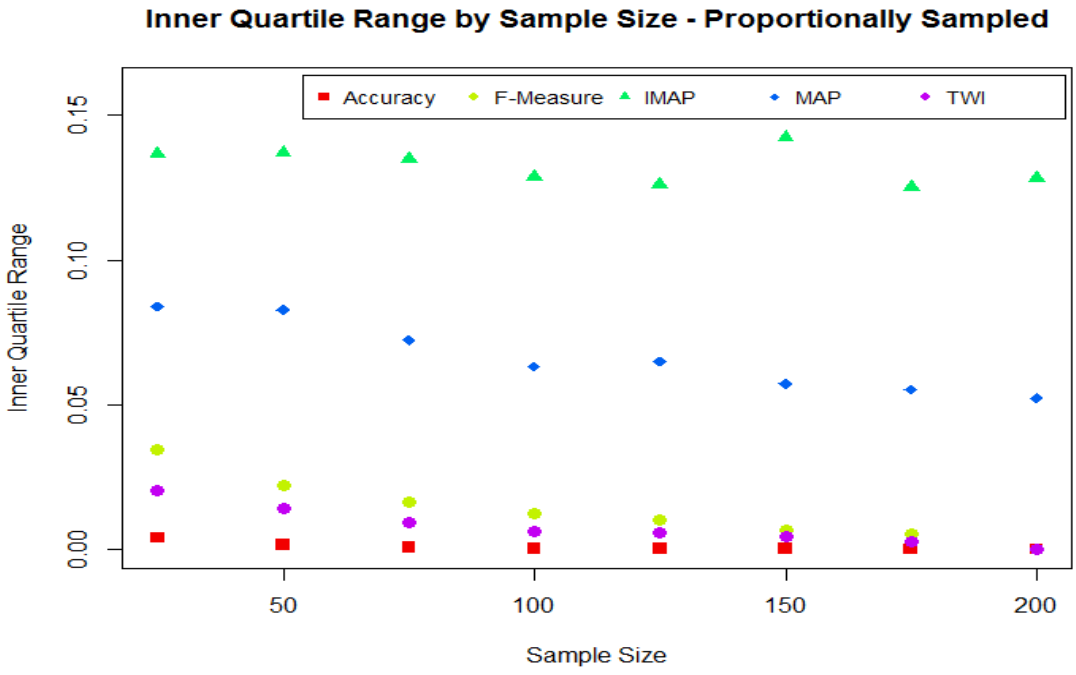


Figure 2

Correlation

The experimental results of the ER runs at every sample size were tested for correlation against the count based measures. The count based measures used in ER were all correlated strongly as expected as were the mean average precision measures used in IR. Likewise, this correlation extended to the ER and IR measures. These correlation results vary with the settings for the two inferred measures. The chart in figure 3 shows the correlation with the optimal settings, a very high match strength score, of each measure with all of the others as well as the significance, indicated with 3 stars for highly significant.

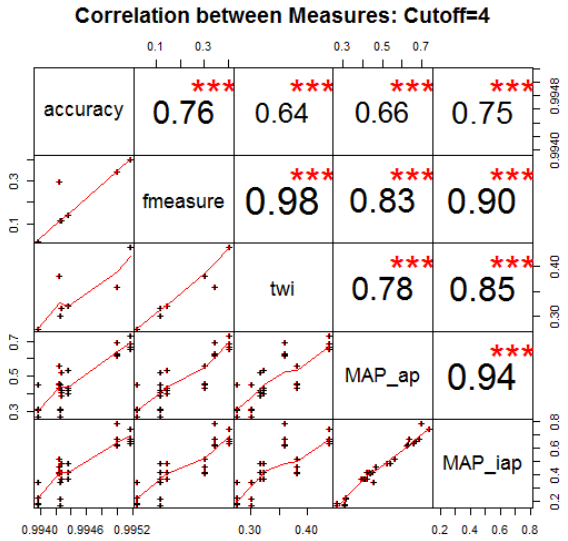


Figure 3

Processing Times

The ER count based measures processed more quickly than the IR measures. The typical maximum for the ER measures was .15 seconds while for the IR measures it was 4 seconds. The minimum times, those for the smallest samples, were instantaneous for the ER measures and 1 second for the IR measures.

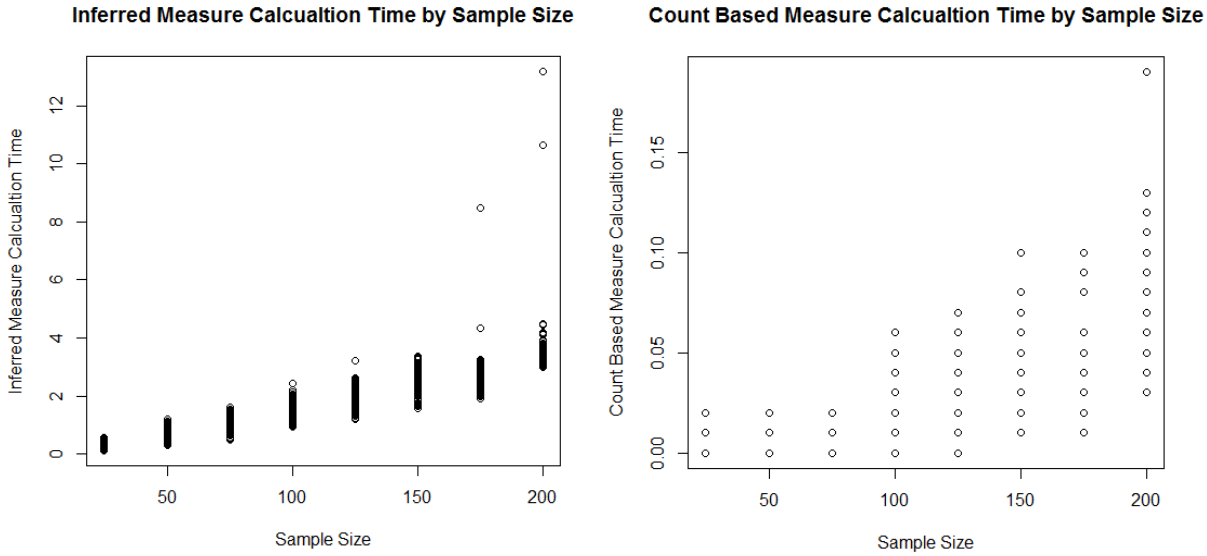


Figure 4

Variance

In figure 5 at a sample size of 25, the smallest sample of clusters collected, all of the measures have their largest variance values. It is clear from this figure that the count based measures have a much smaller variance than the inferred measures even when the sample size is small. At a sample size of 175 clusters in figure 5, the count based measures are converging as expected on the population parameter value. Conversely, the inferred measures maintain a similar distribution regardless of the sample size.

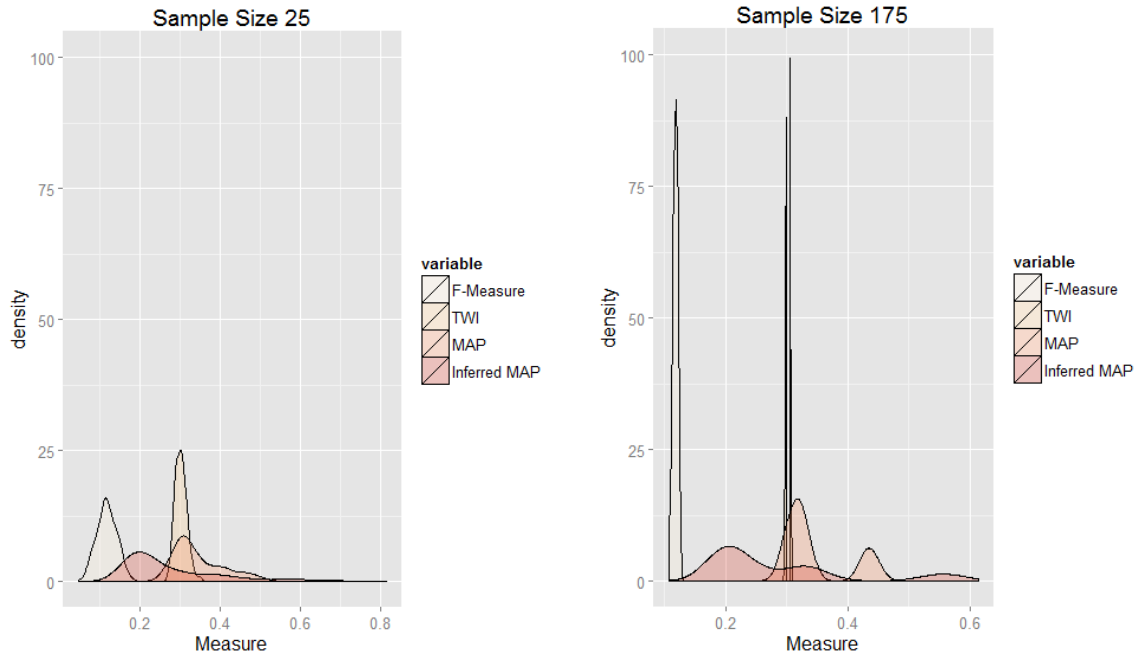


Figure 5

DISCUSSION

Doing ER is needed to combine datasets for healthcare, research and business. Knowing that the ER produced quality results is essential in order for the results to have value. Determining the error rate of ER results is very difficult unless the truth is already known, for the whole or at least a subset of the matched population. Inference is the only efficient solution. Datasets are increasing in size to such an extent that without a scalable solution the quality of ER results will be unknown. To address this problem inferred error rates have been tested in this study to determine their efficacy.

Samples were taken at random and proportionally. There was no difference produced by the different sampling methods. However, the results did show that the adapted MAP and infMAP were highly correlated with the standard count based measures used in ER suggesting that they could be a substitute for a count based evaluation. The count based measures have less variance so are more accurate at smaller samples given a benchmark. A problematic aspect of the adapted IR measures was that processing time was much slower compared with the ER methods at all sample sizes. It should be considered that a truth set is required for the count based measures but not the inferred. The efficiency of ER evaluation is paramount as dataset sizes increase. Runtime increased with sample size for both count based and inferred

measures and correlated results of sample sizes were positive for the both the inferred measures and the count based measures.

LIMITATIONS

The experiments in this study were done using synthetic data. Since the data was generated using real identities it is unlikely that there would be differences if real data was used but this will need to be tested. Further, the synthetic set size is small compared to many real data and the subset used for the experiments was very small compared with many datasets in healthcare today. The slow processing time of the IR measures versus the ER measures might be improved with a larger set. Processing time was not tested at very large sample sizes which is where an inferred measure is the most valuable and effective.

CONCLUSION

The goal of this research was to empirically test sampling for assessing ER error rates and develop an inferred method of ER evaluation based on the inferred measures used in IR. The key contribution of this research has been a proof of concept in a controlled environment that demonstrates that sampling without a known truth is effective for assessing ER error rates. Further, these tests have demonstrated that sampling can be successfully be used with count based as well as alternative measures. Previously all past reliable ER assessments were descriptive, so a measure was only considered reliable if it was descriptive. I have provided examples of inferred ER assessment measures that are reliable. Therefore, reliable ER assessments can be inferred. This approach to ER error rate evaluation is both novel and significant. Sampling has not been studied in the current ER literature and is going to be a necessary addition in our future. This research has been a proof of concept for sampling in ER and has demonstrated that sampling can be successfully used for both count based as well as the newly adapted IR measures.

The process for using the adapted measures is listed in the following steps:

- After completing ER, randomly sample .5% of the pairs.

  This is calculated by taking the average number of pairs per cluster multiplied by the total number of clusters multiplied by .05 and finally dividing by the average number of pairs per cluster.

- Rank the cluster pairs based on their similarity score and choose the scores that will be the breakoff points for pseudo true and false as well as unknown. Ranking based on similarity is almost as old as the ER field and has been an accepted method used by researchers since Felleg & Sunter.

- Using these values, compute the inferred average precision.

Accuracy

This is a first step into sampling and its effects on ER measures. The inferred measures presented here have weakness since the variance is large as compared with the variance of count based measures. Variance will typically get narrower as the sample size is increased. When considering the variance of an ER measure the optimal outcome would be a relatively narrow variance for a sampled set. In the case of

the count based measures the variance behaved as expected and steadily decreased as the sample size was increased. The inferred measures; however, had a much smaller decrease in variance as the sample size grew and started out with a larger variance. They do also have strengths since the results do not require the existence of a benchmark or truth set. Since these test sets are very small compared with real world sets we can assume the time and effort required to complete a sample would be considerably larger but that evaluating the whole ER set would be impossible. Given that sampling is the only option for large datasets a sampled measure that is independent of previous results would be valuable.

Efficiency

To determine relative efficiency, the inferred measures were compared with existing count based measures in setup time and runtime. The time needed to process each sample within each run was captured along with the measure scores. The outcome was a longer processing time for IR measures than ER measures but only if a truth or benchmark set is available. Obtaining and processing this set is very time consuming in itself. This suggests there is a no real processing time advantage for IR or ER.

Scalability

The adapted IR measures were compared with existing count based measures using different size data sets to determine if there was a relationship. The samples worked well at smaller sizes which indicates that a sample can be used to determine the quality of the whole. In both count based measures and inferred measures the samples were correlated with the total known outcome. The advantage of the inferred measures in scalability is their independence from other results meaning that these measures could be applied to large data sets without previous knowledge or results potentially providing a large scalability advantage.

REFERENCES

Bean, M. A. (2001). Probability: The Science of Uncertainty: with Applications to Investments, Insurance, and Engineering. American Mathematical Society.

Christen, P. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Berlin: Springer.

Dunn, H. L. (1946). Record Linkage. Joint Conference of the Vital Statistics Council for Canada and the Dominion Council of Health. Ottawa, Ontario.

Felleg, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 1183-1210.

Goutte, C., & Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation. Proceedings of the European Colloquium on IR Resarch (ECIR'05) (pp. 345-359). Meylan, France: Springer.

Hirsch, W. (2009). Information Retrieval, A Health and Biomedical Perspective. New York: Springer-Verlag New York.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). An Introduction to Information Retrieval. Cambridge, England: Cambridge University Press.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959, October 16). Automatic Linkage of Vital Records. Science, pp. 954-959.

Talburt, J. (2010). In Entity Resolution and Information Quality. Morgan Kaufmann.

Talburt, J. R., & Zhou, Y. (2015). Entity Information Life Cycle for Big Data: Master Data Management and Information Integration. Waltham, MA: Elsevier Science.

Voorhees, E. M., & Hersh, W. (2012). Overview of the TREC 2012 Medical Records Track. The Twenty-First Text REtrieval Conference (TREC 2012) Proceedings. National Institute of Standards and Technology.

Winkler, W. E. (2007). Automatically Estimating Record Linkage False Match Rates. Washington, DC: U.S. Census Bureau, Statistical Research Division.

Yilmaz, E., & Aslam, J. A. (2006). Estimating average precision with incomplete and imperfect judgments. Proceedings of the Fifteenth ACM International Conference on Information and Knowledge Management. ACM Press.

Yilmaz, E., Kanoulas, E., & Aslam, J. A. (2008). A Simple and Efficient Sampling Method for Estimating AP and NDCG. Proceedings of the Thirty-First Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008), (pp. 603–610). Singapore.