Record Linkage Measures in an Entity Centric World

Matthew Michelson MMICHELSON@FETCH.COM
Sofus A. Macskassy SOFMAC@FETCH.COM
Fetch Technologies, 841 Apollo St, Ste. 400, El Segundo, CA 90245 USA

Abstract

For unsupervised clustering, traditional accuracy metrics based on the constituent records do not often reflect the accuracy at the cluster level. For a specific example, consider entity resolution where the goal is to cluster records across multiple, heterogeneous data sources into "entities." Measuring the accuracy of entity resolution is not as simple as applying the well known record level metrics of precision and recall. Rather than using traditional tuple-based metrics for accuracy, we posit that new, entity-based metrics should be defined instead. Defining entity-level metrics gains users a less source biased, yet deeper insight into entity resolution performance. We show that traditional record linkage metrics are not appropriate, and offer some early thoughts on entity-centric measurements that are more so.

1. Introduction

Unsupervised clustering of records, without knowing the target clusters (or how many there should be) is a challenging flavor of the clustering problem (Pelleg & Moore, 2000). One practical variant of such a clustering is *entity resolution* where the goal is to group records from multiple sources into entities. Entity resolution is a generalization of the *record linkage* problem of finding matching records across just two structured sources. While record linkage is concerned with matching the records, entity resolution is focused on clustering matching records across sources into higher level entities. Although the metrics for measuring record linkage are well established in the community, these metrics are inappropriate for the entity resolution task, as we show. Instead of record-centric measures, users should employ entity-centric measures.

Although the record linkage problem has been around for a while (Fellegi & Sunter, 1969), it still receives attention (Minton et al., 2005; Bilenko & Mooney, 2003), which is a tribute to both its difficulty and applicability. As an example, consider the records shown in

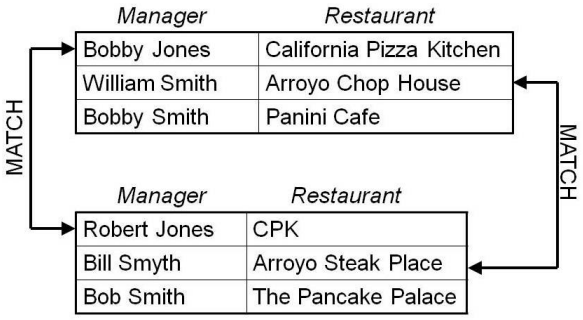

Figure 1. Record linkage

Figure 1. Here the goal is to match the records about restaurant managers, each of which has a field with the manager's name and the restaurant he/she manages. The difficulties arise because of differences in the attribute values, etc.

As stated above, entity resolution generalizes from just the two sources of structured data in record linkage, to multiple, heterogeneous sources of information. Sources can include mentions in natural language text, traditional database records, or even short terms such as user provided "tags," as shown in Figure 2. In this figure, the restaurant manager data comes from structured sources, news text, images, etc. across multiple sources. Rather than just finding the records to link across two structured sources, as in record linkage, entity resolution merges sets of records together into a cluster that represents the "entity" described in the data. In this case, we merge into "restaurant manager" entities.

As machine learning scales to the Web, entity resolution has received more attention as a direct application of using machine learning to make sense of the information overload by finding the entities of interest across the disparate information. For instance, while it is difficult to use record linkage to track entities over time, an entity centric view does allow such analysis. Viewing data at an entity-centric level is much more useful in the context of disparate sources.

However, although there has been a lot of attention to this problem, even very recently (e.g. (Benjelloun et al., 2009; Singla & Domingos, 2006; Bhattacharya

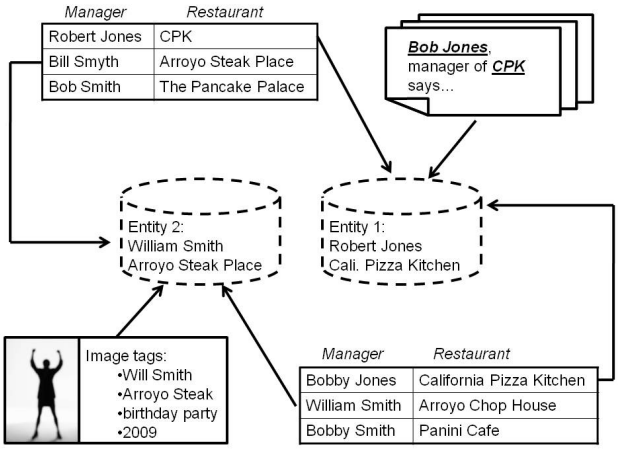

Figure 2. Entity Resolution

& Getoor, 2007)), one area that lags behind is a consistent measurement for the accuracy of entity resolution. Given the similarities between entity resolution and record linkage, it may seem appropriate to apply record linkage metrics to the entity resolution problem. However, an entity-centric view of results is quite different from the record-level view. As we posit in this paper, such record-level analysis does not provide sufficient entity-level information needed to determine the utility of an entity resolution system. Instead, we propose measures that are more focused on measuring the entities themselves (as clusters), rather than the matching records.

The rest of this paper is organized as follows. Section 2 describes traditional record linkage metrics and how they are inappropriate for measuring the entity resolution task. Section 3 outlines metrics that fit the entity resolution problem better. Section 4 presents our conclusions and argues for future directions.

2. Record Linkage Metrics

Record linkage tasks traditionally use two measures for accuracy, *precision* and *recall*, both of which rely on comparing against a truth table. Precision is defined as the number of correct matches made, out of all of the matches made. Meanwhile, recall is the correct number of matches made, out of the total matches that should have been made. Figure 3 shows a truth table and the matches made by the system. Here, the system made two correct matches out of the three it made, so the precision is 66.66%. The system also missed two of the matches it should have gotten, so the recall is 50%.

While precision and recall are appropriate measures for record-level matching, in the process of entity-resolution they can obfuscate the fact that although records might be matching correctly, they are performing poorly at merging entities. Since entities are es-

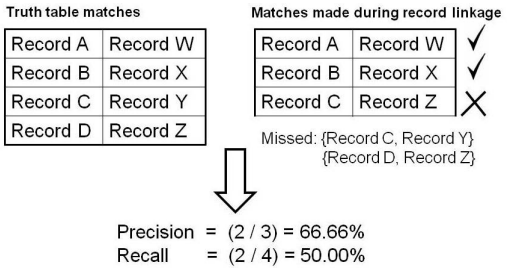

Figure 3. Record Linkage Metrics

sentially clusters, consider the simple case of merging entities based on the transitive closure of matching records. Now, consider the situation of Figure 4. In this figure, we have two sources, with two matches. Now, during record linkage, the system makes the error of matching Record B with Record X, thereby linking all of the records together transitively.

This, in turn, creates a single entity. Now, while the record level metrics shows perfect recall and decent precision at the record level, at the entity level, only a single entity is built. Determining the precision and recall at this entity level already begins to get fuzzy. Recall is fairly analogous. We can argue that the recall is 50% because we did group together all of the records to form one of the entities (either that with records {A, X} or that with {B, Y}). So, we did retrieve one of the entities that we should have.

Regarding precision at the entity level, however, understanding begins to degrade. We could define entity level precision as a measure of how "dirty" the entities are. For instance, we can say that half of the one formed entity is "dirty" because it's composing records are half incorrect and half correct (depending on whether it's considered as Entity 1 or 2). However, across all entities, should this be an average? Further, this is a metric independent of entity size. That is, an entity with 1/2 incorrect records has the same entity-precision as one with 1M / 2M incorrect records. Yet, the second entity is likely to be more problematic given that the 1M incorrect records likely belong to many errant entities, while the 1 incorrect record can only be a single different entity. Therefore, we need a measure that is not size invariant at the entity scale. We address some of these issues with our proposed entity-level metrics of Section 3. Beyond the difficulty in defining the metrics, the record level metrics of precision of 66% and recall of 100% drastically overstate the performance in terms of creating entities (where the argued precision and recall are 50%). This could also happen in reverse. One can imagine the reverse situation, where the entity results, as described above, overstate the record level performance. E.g. in Figure 4, if the system matches Record A and X, and Records A and Y, then an entity {A,X,Y} is formed. At the entity level recall is 50% and precision is 66%, but the record level metrics are both 50%

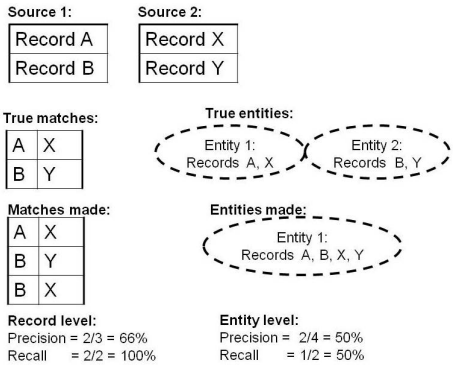

Figure 4. Misleading Record Linkage Metrics

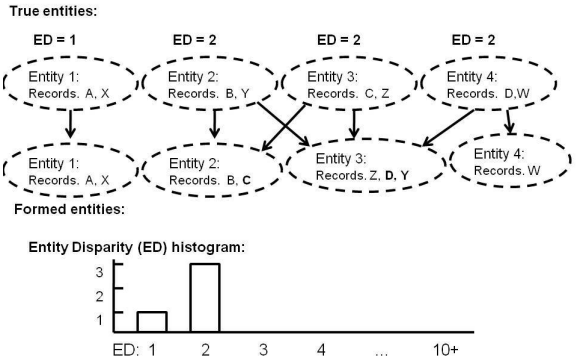

Figure 5. Entity Distribution Histogram

respectively. So there is a disconnect between using record level metrics for entity level data.

Even more problematic, there are important entity level phenomena that are not identifiable when using record level precision and recall. One of the most dangerous phenomena in entity resolution is the "black hole" entity. This is an entity that begins to pull an inordinate amount of records from an increasing number of different true entities into it as it is formed. This is dangerous, because it will then erroneously "match" on more and more records, escalating the problem. However, as shown in Figure 4, precision and recall at the record level can be misleading in the face of a formed black hole.

3. Entity Resolution Metrics

Instead of the traditional precision and recall metrics for record linkage, we propose entity focused metrics instead that are more appropriate for measuring the entity resolution process. These metrics all shift focus to entity resolution as a clustering problem, rather than a record matching problem.

First, we consider is recall at the entity level. That is, given the correct entities, how well did the system cover those entities. In the standard definition, entity recall is (# of correctly formed entities) / (# of known entities). However, this is misleading because it is unclear what a "correctly formed entity" is. For example, records from one true entity can be grouped into multiple entity clusters—which, if any, of the clusters would be "correct"? Or, a cluster could be "mostly correct" in that it contains all records from a true entity but also one record from another entity. Is that a correctly formed entity cluster? Depending on how you count, recall could be arbitrarily high or low.

We therefore propose the *entity distribution* (ED) metric. We define entity distribution as the number of entity clusters a true entity participates in (distributes to). For example, consider Figure 5. In this case, the ED value for Entity 1 is 1 because the entity cluster contains the complete set of records clustered together. Entities 2, 3 and 4, however, have ED values of 2 because their records are each distributed across 2 entity clusters.

In order to combine ED scores, we construct a histogram where the x-axis is the ED score, and the y-axis is the number of entities with that ED score. Figure 5 shows how the given entity resolution yields the ED histogram. The more entities that are formed with low ED values (ideally, 1), the better the entity resolution cleanly constructs the entities.

Things are equally fuzzy for precision. One could consider precision at a per-entity basis. That is, of all the records that form an entity, how many are correct? However, this is plagued with a number of problems: how to combine this score across entities (e.g. average?), how to make it size invariant, how to utilize it for black hole analysis, etc.

Instead, we propose the *entity composition* (EC) metric. We define entity composition as the number of true entities that help form an entity cluster. Given known true entities, for each formed entity cluster we measure how many true entities participate in forming that entity cluster based on their records. To combine EC scores, we construct a histogram similarly to that for ED scores: the x-axis is the EC score, and the y-axis is the number of entities that have that EC score. Figure 6 shows how the given entity resolution yields the EC histogram. The EC histogram is multifaceted. First, it presents a compact summary of entity resolution, at the entity level, since the more cleanly constructed (low EC values) the better the resolution. Second, it allows for analysis of black hole formations. The entities at high EC values are the problematic entities that need to be examined in detail because they are pulling in records from many disparate entities.

For a more detailed analysis, we generate the 3-dimensional graph formed by plotting the EC-histogram, against the entity-cluster purity for each index of the EC histogram. We define entity-cluster purity as the ratio of the number of correct records in an entity cluster over the total number of records in the entity cluster. So, we plot the triple {x, y, z}

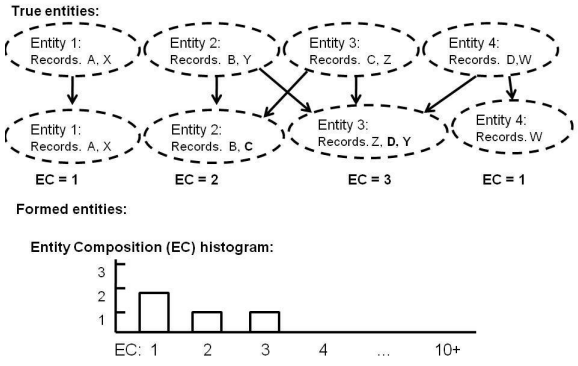

Figure 6. Entity Composition Histogram

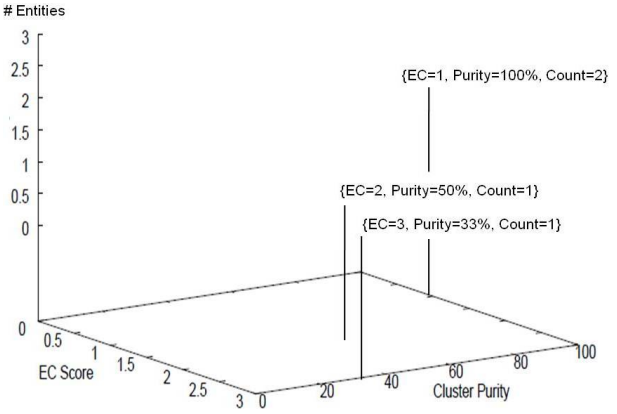

Figure 7. 3-D Entity Composition Histograms

for a given grouping of {EC value, entity-cluster purity, num. entities }. This groups the entities at both their EC value and their level of entity-cluster purity (we bin the entity-cluster purity to make the values discrete). Figure 7 shows an example 3-d plot.

While this detailed plot loses the summarizing power of just the EC histogram, it allows for much more detailed analysis that is dependent on the size of the entity (in terms of how many entities compose that entity), and yields a cleanliness measure for entities of specific EC values. For instance, an entity of EC value 3 with entity precision of 97% is much different than that of EC value 3 with entity precision of 3%. Yet, now such a distinction can be identified.

4. Conclusions and Future Directions

In this paper we presented alternative measures for entity resolution based on entity level metrics rather than record level metrics. Our first metric, a histogram of entity distribution values, visualizes how well entities are reconstructed, somewhat analogous to recall. The second measure, a histogram of entity composition values, presents a mechanism to view how cleanly entities are forming, and is somewhat analogous to precision. A more definite analogue to precision is the 3-d plot of EC histograms versus entity precision, although this plot loses some of the summarization capability of the EC histogram. These measures provide more accurate insight into the entity resolution task than the traditional, record linkage statistics of precision and recall.

The key insight, beyond considering entity level measures, is that using histograms is important because accuracies at the entity level are multi-dimensional. It is not only important to examine how accurate entity resolution forms the entities, but also the frequencies at which it does so, such that seemingly small accuracy problems (such as merging a few records into a single entity) are not undervalued (since this could be a black hole swallowing all records). We plan to run empirical tests on real world data sets to demonstrate that in fact, the record level results are not appropriate and that measures such as those presented here should be used instead.

Finally, we believe such statistics can also increase the understanding of general unsupervised clustering techniques, such as X-Means (Pelleg & Moore, 2000), which do not know the target clusters a priori. Since the entity resolution problem is a specific version of the clustering problem, these metrics will generalize to clustering. In fact, in many cases there are specific clustering issues that current metrics do not uncover, such as the black-hold phenomena, for which our method is well suited.

Acknowledgements

This work was sponsored in part by the Office of Naval Research under award number N00014-07-C-0923. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Office of Naval Research or the U.S. Government.

References

Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S. E., & Widom, J. (2009). Swoosh: a generic approach to entity resolution. *VLDB J.*, *18*, 255–276.

Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Trans. Knowl. Discov. Data*, *1*, 5.

Bilenko, M., & Mooney, R. J. (2003). Adaptive duplicate detection using learnable string similarity measures. *Proceedings of ACM SIGKDD* (pp. 39 – 48).

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*, 1183–1210.

Minton, S. N., Nanjo, C., Knoblock, C. A., Michalowski, M., & Michelson, M. (2005). A heterogeneous field matching method for record linkage. *Proceedings of ICDM* (pp. 314–321).

Pelleg, D., & Moore, A. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. *Proceedings of ICML* (pp. 727–734).

Singla, P., & Domingos, P. (2006). Entity resolution with markov logic. *Proceedings of ICDM*.