Joint Deduplication of Multiple Record Types in Relational Data

Aron Culotta, Andrew McCallum
University of Massachusetts
140 Governor's Drive
Amherst, MA USA
{culotta,mccallum}@cs.umass.edu

ABSTRACT

Record deduplication is the task of merging database records that refer to the same underlying entity. In relational databases, accurate deduplication for records of one type is often dependent on the decisions made for records of other types. Whereas nearly all previous approaches have merged records of different types independently, this work models these inter-dependencies explicitly to collectively deduplicate records of multiple types. We construct a conditional random field model of deduplication that captures these relational dependencies, and then employ a novel relational partitioning algorithm to jointly deduplicate records. For two citation matching datasets, we show that collectively deduplicating paper and venue records results in up to a 30% error reduction in venue deduplication, and up to a 20% error reduction in paper deduplication.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*clustering*

General Terms: Algorithms, Performance

Keywords: record linkage, deduplication, conditional random fields

1. INTRODUCTION

A common prerequisite for knowledge discovery is accurately combining data from multiple, heterogeneous sources into a unified, mineable database. An important step in creating such a database is *record deduplication*: consolidating multiple records that refer to the same abstract entity.

Most historical approaches have framed the deduplication problem as a set of independent, pairwise decisions. More recently, McCallum and Wellner [5] and Parag and Domingos [6] have demonstrated that making multiple deduplication decisions collectively can provide better results than historical approaches. These models are types of conditional random fields (CRFs) [3], where the observed nodes are mentions, and the predicted nodes are the deduplication decisions for each pair of nodes. The models are "collective" in the sense that mentions are clustered based not only on their distance to each other, but also on their distance to all other mentions. By treating deduplication decisions in *dependent relation* to each other, inconsistencies and noise in the similarity metric may be overcome.

We extend this work to the case of relational databases, where *the identity of a record often depends on the identities of related records*. For example, consider a database of research papers, where records can be of type *paper* or *venue*. If two *paper* records are labeled as duplicates, then it follows that the *venue* records corresponding to those papers should also be labeled as duplicates. The converse is more subtly true: if two *venues* are duplicates, then this may slightly increase the probability that their corresponding *papers* are duplicates. We propose a CRF model that leverages these subtle interdependencies to make deduplication decisions collectively across multiple record types, and we validate its performance on two real-world datasets.

2. MODEL

The model is an instance of a conditional random field that jointly models the conditional probability of multiple deduplication decisions given an observed relational database.

CRFs [3] are undirected graphical models encoding the conditional probability of a set of output variables $\mathbf{Y}$ given a set of evidence variables $\mathbf{X}$. Let $\mathbf{X}$ be a collection of random variables representing observed record mentions in a database. For clarity, assume there are only two types of records, $\mathbf{X} = (\mathbf{X}^a, \mathbf{X}^b)$, where $\mathbf{X}^a = (X_1^a, \ldots, X_n^a)$, $\mathbf{X}^b = (X_1^b, \ldots, X_m^b)$. The goal of deduplication is to partition $\mathbf{X}$ into clusters of records that refer to the same abstract entity.

To this end, we define a collection of binary random variables $\mathbf{Y} = (\mathbf{Y}^a, \mathbf{Y}^b)$ indicating whether or not two records are duplicates. For example, Y_{ij}^a indicates whether or not records X_i^a and X_j^a are coreferent. We also define the binary random variables $\mathbf{R}$, where R_{ij}^{ab} indicates whether some relation R holds between record mentions X_i^a and X_j^b.

For example, in a research paper database, $\mathbf{X}^a$ represents the set of paper records, $\mathbf{X}^b$ represents the set venue records, Y_{ij}^a indicates whether X_i^a and X_j^a are duplicates, and R_{ij}^{ab} indicates whether paper X_i^a was published at venue X_j^b. We model the conditional distribution $P(\mathbf{Y}^a, \mathbf{Y}^b | \mathbf{X}, \mathbf{R})$.

Let $\mathbf{x}_{ij}^{ab} = \langle x_i^a, x_j^a, x_i^b, x_j^b \rangle$ be a pair of observed paper records and their corresponding venue records. To capture

CIKM'05, October 31–November 5, 2005, Bremen, Germany.
ACM 1-59593-140-6/05/0010.

the dependence between y^a_{ij} and y^b_{ij}, we factorize the potential functions to consider them jointly, resulting in:

$$p(\mathbf{y}^a, \mathbf{y}^b|\mathbf{x}, \mathbf{r}) = \frac{1}{Z_\mathbf{x}} \exp\Big(\sum_{i,j,l} \lambda_l f_l(\mathbf{x}^{ab}_{ij}, y^a_{ij}, y^b_{ij}, r^{ab}_{ij}) + \sum_{i,j,k} \lambda_* f_*(y^a_{ij}, y^a_{jk}, y^a_{ik}, y^b_{ij}, y^b_{jk}, y^b_{ik})\Big)$$

where f_l are feature functions, f_* are consistency checking functions used to enforce transitivity among deduplication decisions, and $Z_\mathbf{x}$ is a normalizer. The parameters λ are learned by maximizing a product of local marginals [5, 4].

MAP inference in this model corresponds to finding the solution to $\mathbf{y}^* = (\mathbf{y}^{a*}, \mathbf{y}^{b*}) = \text{argmax}_\mathbf{y}\, p_\Lambda(\mathbf{y}^a, \mathbf{y}^b|\mathbf{x}^a, \mathbf{x}^b, \mathbf{r})$ that is, finding the most probable deduplication decisions $\mathbf{y}^*$ given $\mathbf{x}^a, \mathbf{x}^b, \mathbf{r}$ and the learned parameters Λ.

Because exact inference here is intractable, we follow recent work which finds an equivalence between graph partitioning algorithms and inference in certain undirected graphical models [1, 5]. We first transform our graph to a weighted, undirected graph that only contains vertices for variables $\mathbf{x}$ and has edges weighted by the (log) clique potential for each pair of vertices. The value on these edges depends on which type of records they join.

For paper edges, we define the weight

$$w^a_{ij} = \sum_{y^b_{ij} \in \{0,1\}} \Big(\sum_l \lambda_l f_l(\mathbf{x}^{ab}_{ij}, y^a_{ij} = 1, y^b_{ij}, r^{ab}_{ij}) - \sum_l \lambda_l f_l(\mathbf{x}^{ab}_{ij}, y^a_{ij} = 0, y^b_{ij}, r^{ab}_{ij})\Big)$$

and similar weights w^b_{ij} for venue edges: It can be shown that the optimal partitioning of this graph corresponds to the optimal configuration $\mathbf{y}^*$ in the original undirected graphical model. Here, the number of partitions is unknown, as it corresponds to the number of unique records.

Because traditional partitioning algorithms do not account for the known relations between clusters that exist in our data, we develop a novel, *relational agglomerative* clustering algorithm that exploits these dependencies. This algorithm iteratively merges nodes, enforcing relations and adjusting weights accordingly. For more details, we refer the reader to our technical report [2]

3. EXPERIMENTS

We experiment on two datasets of research paper citations: **Citeseer** (1500 citations) and **Cora**[1], (1800 citations). Feature functions include string matches and cosine similarity of citation fields.

Table 1 shows the pairwise F1 performance of two systems: **joint** is the system we have advocated in this paper, and **indep** is the system which deduplicates records of different types *independently*, although records of the *same* type are deduplicated collectively as in McCallum and Wellner [5].

Venue performance improves considerably in the joint model, which is reasonable considering the strong influence paper deduplication has on venue deduplication. The joint model obtains a 5% absolute recall boost in **Citeseer**, and a 9% boost in **Cora** data. This is because the hard constraint requiring the venues of duplicate papers to be merged often correctly merges venues with dissimilar surface forms.

[1] http://www.cs.umass.edu/~mccallum/data/cora-refs.tar.gz

		Paper		Venue	
		indep	**joint**	**indep**	**joint**
Citeseer	**constraint**	88.9	**91.0**	79.4	**94.1**
	reinforce	92.2	92.2	56.5	**60.1**
	face	88.2	**93.7**	80.9	**82.8**
	reason	**97.4**	97.0	75.6	**79.5**
	Micro Avg.	91.7	**93.4**	73.1	**79.1**
Cora	**kibl**	92.9	**93.3**	93.6	**99.3**
	fahl	**95.5**	95.0	87.3	**99.7**
	utgo	79.9	**84.0**	51.7	**60.4**
	Micro Avg.	89.4	**90.8**	77.5	**84.5**

Table 1: Pairwise F1 deduplication performance.

More interestingly, a noticeable improvement in paper deduplication is attained by the collective model. Part of this is due to the precision enhancement provided by the clustering algorithm. Workshop and technical report versions of journal or conference papers with the same title are correctly not merged when the venues are accurately identified. Also, error analysis suggests that papers that would not have been otherwise merged were merged because their venues were determined to be coreferent.

4. CONCLUSIONS

We have introduced a collective model for deduplication of relational data and empirically demonstrated its advantage over competing methods. Future work includes modeling data where the relations $\mathbf{R}$ are unknown, for example discovering *AdvisorOf* relations between authors.

5. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, by U.S. Government contract #NBCH040171 through a subcontract with BBNT Solutions LLC, by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0326249, and by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010. Any opinions, findings, conclusions or recommendations expressed are the author(s)' and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *In IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(11):1222–1239, 2001.

[2] A. Culotta and A. McCallum. A conditional model of deduplication for multi-type relational data. Technical Report IR-443, Center for Intelligent Information Retrieval, University of Massachusetts, 2005.

[3] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

[4] A. McCallum and C. Sutton. Piecewise training with parameter independence diagrams: Comparing globally- and locally-trained linear-chain crfs. In *NIPS 2004 Workshop on Learning with Structured Outputs*, 2004.

[5] A. McCallum and B. Wellner. Conditional models of identity uncertainty with application to noun coreference. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA, 2005.

[6] Parag and P. Domingos. Multi-relational record linkage. In *Proceedings of the KDD-2004 Workshop on Multi-Relational Data Mining*, pages 31–48, Aug. 2004.