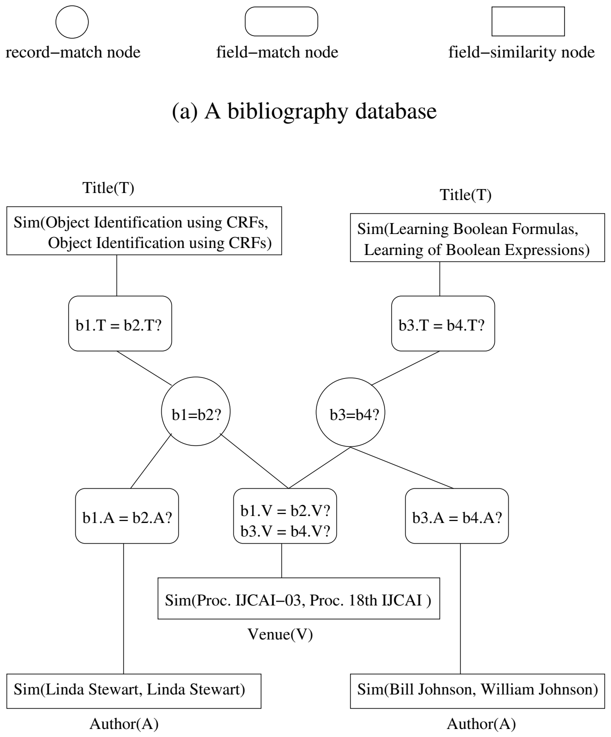
Collective Object Identification

Parag Singla **Pedro Domingos**
Department of Computer Science and Engineering
University of Washington
Seattle, WA 98195-2350, U.S.A.
{*parag, pedrod*}@*cs.washington.edu*

In many domains, the objects of interest are not uniquely identified, and the problem arises of determining which observations correspond to the same object. For example, in information extraction and NLP we need to determine which noun phrases refer to the same entity. When merging multiple databases, a problem of keen interest to many large scientific projects, businesses, and government agencies, we need to determine which records represent the same entity and should therefore be merged. This problem, first placed on a firm statistical footing by Fellegi and Sunter [1969], is known by the name of object identification, record linkage, de-duplication and others. Most approaches described to solve this problem are variants of the original Fellegi-Sunter model, in which object identification is viewed as a classification problem: given a vector of similarity scores between the attributes of two observations, classify it as "Match" or "Non-match." A separate match decision is made for each candidate pair, followed by transitive closure to eliminate inconsistencies. Typically, a logistic regression model is used. We call this the standard model.

Record	Title	Author	Venue
b1	Object Identification using CRFs	Linda Stewart	Proc. AAAI−05
b2	Object Identification using CRFs	Linda Stewart	Proc. 20th AAAI
b3	Learning Boolean Formulas	Bill Johnson	Proc. AAAI−05
b4	Learning of Boolean Expressions	William Johnson	Proc. 20th AAAI


record−match node
field−match node
field−similarity node
Title(T)
Sim(Object Identification using CRFs, Object Identification using CRFs)
Title(T)
Sim(Learning Boolean Formulas, Learning of Boolean Expressions)
b1.T = b2.T?
b3.T = b4.T?
b1=b2?
b3=b4?
b1.A = b2.A?
b1.V = b2.V?
b3.V = b4.V?
b3.A = b4.A?
Sim(Proc. IJCAI−03, Proc. 18th IJCAI)
Venue(V)
Sim(Linda Stewart, Linda Stewart)
Author(A)
Sim(Bill Johnson, William Johnson)
Author(A)

(a) A bibliography database

(b) Collective model (fragment)

Figure 1: Example of collective object identification. For clarity, we have omitted the edges linking the record-match nodes to the corresponding field-similarity nodes.

Making match decisions separately ignores that information gleaned from one match decision may be useful in others. For example, if we find that a paper appearing in *Proc. IJCAI-03* is the same as a paper appearing in *Proc. 18th IJCAI*, this implies that these two strings refer to the same venue, which in turn can help match other pairs of IJCAI papers. In this paper, we propose an approach which accomplishes this propagation of information. Our approach makes decisions collectively, performing simultaneous inference for all candidate match pairs, and allowing information to propagate from one candidate match to another via the attributes (or fields) they have in common. Our model is based on conditional random fields [Lafferty *et al.*, 2001]. We call our model the collective model. Figure 1(a) shows a four-record bibliography database and 1(b) shows the corresponding graphical representation for the candidate pairs (b_1, b_2) and (b_3, b_4) in the collective model. There are three types of nodes in the figure. *Record-match nodes* are Boolean-valued and they correspond to asking the question "Is record b_i the same as record b_j?" *Field-match nodes* are also Boolean-valued and they correspond to asking the question "Do field values $b_i.F$ and $b_j.F$, for the field F, represent the same underlying property?" *Field-similarity nodes* are real-valued nodes taking values in the domain [0, 1] and they encode how similar two field values are, according to a pre-defined similarity measure. The values of these nodes can be directly computed from data, and hence they are also called the *evidence nodes*. Intuitively, an edge between two nodes represents the fact their values directly influence each other. Note how dependencies flow through the shared field-match node corresponding to the venue field. Inferring that b_1 and b_2 refer to the same underlying paper will lead to the inference that the corresponding venue strings “Proc. IJCAI-03” and “Proc. 18th IJCAI” refer to the same underlying venue, which in turn might provide sufficient evidence to merge b_3 and b_4. In general, our model can capture complex interactions between candidate pair decisions, potentially leading to better object identification.

For random fields where maximum clique size is two and all non-evidence nodes are Boolean, the inference problem can be reduced to a graph min-cut problem, provided certain constraints on the parameters are satisfied [Greig *et al.*, 1989]. Our formulation of the problem satisfies these constraints. Since min-cut can be solved exactly in polynomial time, we have a polynomial-time exact inference algorithm for our model. We follow the standard approach of gradient descent to learn the parameters. Calculating the exact derivative is intractable as it involves an expectation over an exponential number of configurations. We use a voted perceptron algorithm [Collins, 2002], which approximates this expectation by the feature counts of the most likely configuration, which we find using our polynomial-time inference algorithm with the current parameters.

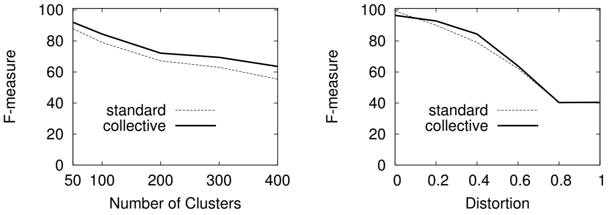
Combining models is often a simple way to improve accuracy. We combine the standard and collective models using logistic regression. For each record-match node in the training set, we form a data point with the outputs of the two models as predictors, and the true value of the node as the response variable. We then apply logistic regression to this dataset. Notice that this still yields a conditional random field.

We performed experiments on real and semi-artificial databases, comparing the performance of (a) the standard Fellegi-Sunter model using logistic regression, (b) the collective model, and (c) the combined model.

The first set of experiments was on Cora database, which is a collection of 1295 different citations to computer science research papers. We cleaned it up by correcting some labels and filling in missing values. This cleaned version contains references to 132 different research papers. We used the author, venue, and title fields. The second set of experiments was done on the BibServ.org database, which is the result of merging citation databases donated by its users, Citeseer, and DBLP. We experimented on the user-donated subset of BibServ, which contains 21,805 citations. Table 1 reports the results for these databases. The combined model gives the best performance on both Cora and BibServ, followed by the collective model. Transitive closure helps these on Cora but hurts all models on BibServ (where recall was close to 100% even without transitive closure). The best combined model outperforms the best standard model in F-measure by about 2% on Cora and 3% on BibServ.

Table 1: F-measures on Cora and BibServ before and after transitive closure.

Model	Cora		BibServ	
	Before	After	Before	After
Standard	86.9%	84.7%	82.7%	68.5%
Collective	87.4%	88.9%	82.8%	73.6%
Combined	85.8%	89.0%	85.6%	76.0%



(a) F-measure vs. number of clusters

(b) F-measure vs. distortion level

Figure 2: Experimental results on semi-artificial data.

To further observe the behavior of the algorithms, we generated variants of the Cora database by taking distinct field values from the original database and randomly combining them to generate distinct papers. Figures 2(a) and 2(b) compare the performance of the collective and standard models as number of clusters and level of distortion in the data are varied, respectively.[1] The collective model clearly dominates the standard model over a broad range of number of clusters and level of distortion.

In summary, determining which observations correspond to the same object is a key problem in information integration, citation matching, natural language, vision, and other areas. We have developed a collective approach to this problem, where information propagates among related decisions via shared field values, and shown experimentally that it outperforms the standard one of making decisions independently.

Acknowledgments

This research was partly supported by ONR grant N00014-02-1-0408, by a gift from the Ford Motor Co., and by a Sloan Fellowship awarded to the second author.

References

[Collins, 2002] M. Collins. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proc. EMNLP-02*, pages 1–8, 2002.

[Fellegi and Sunter, 1969] I. Fellegi and A. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64:1183–1210, 1969.

[Greig *et al.*, 1989] D. M. Greig, B. T. Porteous, and A. H. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society, Series B*, 51:271–279, 1989.

[Lafferty *et al.*, 2001] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th ICML*, pages 282–289, 2001.

[1] For clarity, we have not shown the curves for the combined model, which are similar to the collective one’s.