Felix: Exploiting Specialized Subtasks in Markov Logic Networks for Higher Efficiency and Quality

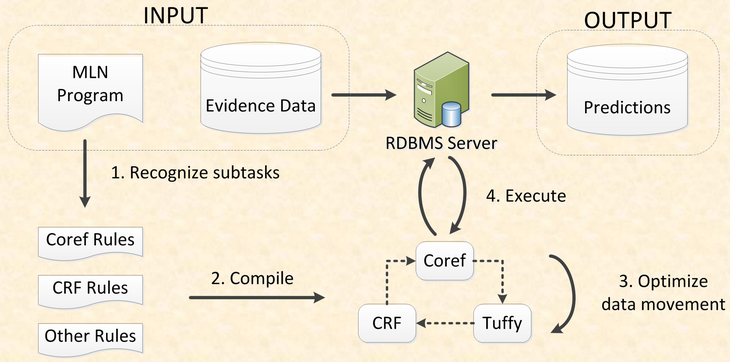
Feng Niu, Ce Zhang, Chris Ré, and Jude Shavlik
University of Wisconsin-Madison

Executive Summary

- Machine Reading (MR) requires joint inference
  - Desirable to have a versatile, easy-to-use language
- Markov logic network (MLN) is such a language
  - But current approach to inference is monolithic: one algorithm for entire program
  - Suboptimal scalability and quality
- Key observation: MLNs in MR may contain subtasks
  - E.g., NER, coreference, link prediction, etc.
- Felix hypothesis: We can get higher scale and quality by exploiting those subtasks
  - Idea: solve subtasks with specialized algorithms
  - Preliminary results show dramatic improvement in both efficiency and quality

Felix at Work

- Felix Approach to MLN Inference
  - 1. Recognize specialized subtasks in an MLN program
  - 2. Compile a logical plan by assigning subtasks to specialized algorithms (aka "statistical operators")
  - 3. Optimize data movement between operators
  - 4. Execute operators according to optimized plan



Specialized Subtasks in Felix

Subtasks recognized by Felix (Step 1)

Logistic Regression

$$P[\text{tag} = t] = \frac{\exp(\sum_f w_{f,t})}{\sum_t \exp(\sum_f w_{f,t})} \qquad w_{\text{fea,tag}} \; F(\text{pos}, \text{fea}) \wedge T(\text{pos}, \text{tag})$$

weights · features · labels

Conditional Random Field

$$P[T = \vec{t}] = \frac{1}{Z(F)} \exp\left\{\sum_{f,i} w_{f,t_i,t_{i-1}}\right\}$$

$w_{\text{fea,tag1,tag2}}$
$F(\text{pos}, \text{fea})$
$\wedge\, T(\text{pos}-1, \text{tag1})$
$\wedge\, T(\text{pos}, \text{tag2})$

Correlation Clustering (Coreference)

∞ $\text{Coref}(x,y),\ \text{Coref}(y,z) \Rightarrow \text{Coref}(x,z)$
-8 $\text{DiffVenue}(x,y) \Rightarrow \text{Coref}(x,y)$
3 $\text{SimTitle}(x,y) \Rightarrow \text{Coref}(x,y)$

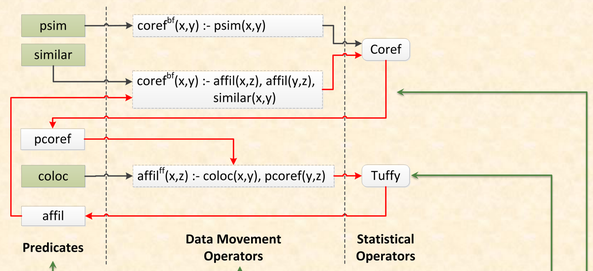
Input: Graph $G = (V, w)$ where $w: V \times V \mapsto \mathbb{R}$
Output: A partition of V
Goal: Minimize weights of violated edges

solid: positive edges
dash: negative edges
red: violated edges

There are specialized algorithms for them (Step 2)
[Hilbe 2009; Lafferty et al. 2001; Arasu et al. 2009]

Example Plan in Felix

- Task: Extract person-org affiliations from webpages
- Subtasks
  - coreference (e.g., pcoref for person coref)
  - link prediction (e.g., affil for affiliations)



Dark boxes = evidence; light boxes = predictions

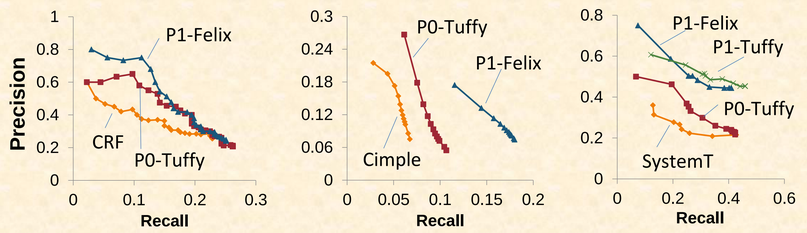
Data manipulation optimized by Felix using a Relational Database Management System (Step 3)

Generic MLN inference handled by Tuffy, our prior work of an MLN inference engine [Niu et al. 2011]

Joint inference via cycles through operators

Experiments

- Validation: Felix achieves higher efficiency and quality than monolithic
- Method: Run Felix, Tuffy, and Alchemy on MLN programs for IE
  - Three tasks: NFL winner-loser, DBLife affiliations, Enron person-phone
  - Baseline extractors: CRF, DBLife Cimple, IBM SystemT
  - Two MLNs each task: P0 without coref subtask; P1 with coref
- Results:
  - Alchemy crashed on all MLN programs
  - Tuffy crashed on P1 of NFL and DBLife
  - Felix scales and achieves best quality
  - Quality: P1 > P0 > baselines



(a) NFL Winner-loser (b) DBLife Affiliation (c) Enron Person-phone

We gratefully acknowledge the support of Air Force Research Laboratory (AFRL) under prime contract no. FA8750-09-C-0181. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA, AFRL, or the US government.

Conclusion & Future Work

- Conclusion
  - Felix achieves higher efficiency and quality by running specialized algorithms for well-studied subtasks buried in MLN programs
- Future Work
  - More specialized subtasks/algorithms, e.g. Cuts
  - More optimization in data movement
  - More systematic recognition/planning
  - Efficient weight learning in Felix (for feedback between operators)

References

- A. Arasu, C. Re, D. Suciu, *Large-scale deduplication with constraints using Dedupalog*, ICDE 2009
- J. Hilbe, *Logistic Regression Models*, Chapman & Hall/CRC Press, 2009
- J. Lafferty, A. McCallum, F. Pereira, *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*, ICML 2001
- F. Niu, C. Re, A. Doan, J. Shavlik, *Tuffy: Scaling up Statistical Inference in Markov Logic Networks using an RDBMS*, VLDB 2011, to appear

Template provided by: "posters4research.com"