Refining Automatically Generated Knowledge Bases in a Scalable, Distributed Fashion

Paul Elliott
University of Oregon
paule@cs.uoregon.edu

David Stevens
University of Oregon
dstevens@cs.uoregon.edu

Today's World Wide Web is a goldmine of text-based information. Much work has been done to integrate this information into a structured *knowledge base*, but this is no easy task. State-of-the-art information extraction systems such as NELL [1], which collates information pulled from the Web into an ontology, are lacking in the way they handle uncertainty. The result is a noisy ontology–a knowledge base laced with uncertain knowledge.

Properly handling the uncertainty within these relational knowledge bases is a classic task for statistical relational learning. In particular, there have been efforts ([2]) to clean up NELL's knowledge base using Markov logic. Techniques such as the Markov Chain Monte Carlo (MCMC) algorithm can help correct uncertainties within the knowledge base by jointly reasoning over the entire dataset. However, NELL's ontology is too large for a simple implementation of MCMC. While work has been done to manually partition how MCMC runs on the knowledge base, this is a time-consuming task and not a scalable solution. A better solution would involve managing the inference task in a distributed fashion, enabling an efficient and scalable solution.

Our project involves leveraging a parallel, distributed belief propagation algorithm to clean up NELL's knowledge base. We propose to leverage GraphLab's Graphical Models toolkit [3] toward this end. Our solution will involve properly partitioning the dataset, running the distributed belief propagation algorithm, and evaluating the resulting ontology comparatively with previous work. Such work presents a much more scalable approach than existing practices.

Time Line

- **Week 3 – 4** Literature review and familiarization with GraphLab
- **Weeks 5 – 7** Design and implementation
- **Weeks 8 – 9** Experimentation and evaluation
- **Week 10** Paper and presentations

References

[1] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. Hruschka Jr, and T. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*, volume 2, pages 3–3, 2010.

[2] N. Lao, T. Mitchell, and W. Cohen. Random walk inference and learning in a large scale knowledge base. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 529–539, 2011.

[3] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. Hellerstein. Graphlab: A new framework for parallel machine learning. *arXiv preprint arXiv:1006.4990*, 2010.