

# Constraint-driven hierarchical information extraction\*

Thomas Lee

University of Pennsylvania, The Wharton School  
Philadelphia, PA 19104  
thomas.lee@wharton.upenn.edu

This abstract summarizes a proposal to use semistructured constraints as a general framework for information extraction. The proposal builds on the intuition that written documents tend to contain a great deal of redundancy: Tell them what you are going to tell them, tell them, tell them what you told them. While current methods overlook or eliminate redundancy, we seek to explicitly represent and exploit both repetition and other natural patterns in semistructured text. We summarize our early research on verification, outline current and future directions, and identify related work.

Our initial work assumes knowledge of both document structure and its attendant constraints for the purpose of extraction verification. Specifically, semantic and structural constraints are used to check wrapper accuracy.

For example, a structural constraint might ensure that the parties to a contract are introduced at the beginning of the document. A foreign-key constraint checks whether the parties are included on a list of approved signers. That signatories are repeated at the end of a contract is represented by an inverse constraint.

A wrapper to extract seven attributes from legal arbitration decisions was applied to 200 cases. Constraint-based verification produced only ten false negatives, the majority of which were due to errors in wrapper delimiters. Further results were reported earlier [Hunter, et al., 2002].

We are currently experimenting with semistructured constraints in hierarchical extraction. Documents with the same semantic and structural constraints are placed in the same class. Class-specific routines fragment the document, extract from each fragment, and recurse on selected components. We propose the use of semistructured constraints as a framework for both fragmentation and extraction.

For example, in addition to simple features like HTML-nesting, constraints can dictate whether certain values or structural properties functionally determine other fragments. A dissenting opinion must correspond to a legal decision with more than one judge.

In future work, we aim to explore induction techniques for learning the constraints used to extract and verify.

We believe that this is the first attempt to systematically exploit semistructured constraints as a framework for extraction and verification. Early work in extraction verification uses supervised techniques to learn statistical characteristics of tuples [Kushmerick, 2000] and attribute domains [Knoblock, et al., 2001]. They can signal errors in a test set but cannot identify the values at fault. Recent work on integration leverages relational associations but not necessarily structural features [Doan, et al., 2003]. Current hierarchical approaches learn basic structural features but not more complex constraints [Califf and Mooney, 1999, Knoblock, et al., 2001, Wang, et al., 1997]

## References

- [Califf and Mooney, 1999] M. E. Califf and R. J. Mooney. Relational learning of pattern-match rules for information extraction. *AAAI-99*.
- [Doan, et al., 2003] A. Doan, P. Domingos and A. Halevy. Learning to match the schemas of databases: A multistrategy approach. *Machine Learning Journal*, 50:279-301.
- [Hunter, et al., 2002] D. Hunter, T. Lee, D. Ong and Y. Yang. Wrapper verification using constraints for semistructured data. *WITS02*.
- [Knoblock, et al., 2001] C. A. Knoblock, K. Lerman, S. Minton and I. Muslea. Accurately and reliably extracting data from the web: A machine learning approach. *IEEE Data Engineering Bulletin*, 23(4):33-41.
- [Kushmerick, 2000] N. Kushmerick. Wrapper verification. *World Wide Web Journal*, 3(2):79-94.
- [Wang, et al., 1997] J. Wang, D. Shasha, G. Chang, L. Relihan, K. Zhang and G. Patel. Structural matching and discovery in document databases. *ACM SIGMOD97*.

---

\* Dan Hunter of the Legal Studies Department, Yingwei Yang, and the University of Pennsylvania Research Foundation's financial support are gratefully acknowledged.