Building Knowledge Bases from the Web

Rajeev Rastogi
Machine Learning, Amazon

Abstract

The web is a vast repository of human knowledge. Extracting structured data from web pages can enable applications like comparison shopping, and lead to improved ranking and rendering of search results. In this talk, I will describe two efforts to extract records from pages at web scale. The first is a wrapper induction system that handles end-to-end extraction tasks from clustering web pages to learning XPath extraction rules to relearning rules when sites change. The system has been deployed in production within Yahoo! to extract more than 500 million records from ~200 web sites. The second effort exploits machine learning models to automatically extract records without human supervision. Specifically, we use Markov Logic Networks (MLNs) to capture content and structural features in a single unified framework, and devise a fast graph-based approach for MLN inference.

Biography

Rajeev Rastogi is the Director of Machine Learning at Amazon. Previously, he was the Vice President of Yahoo! Labs Bangalore, and a Bell Labs Fellow at Bell Labs in Murray Hill, NJ. Rajeev is active in the fields of databases, data mining, and networking, and has served on the program committees of several conferences in these areas. He currently serves on the editorial board of the CACM, and has been an Associate editor for IEEE Transactions on Knowledge and Data Engineering in the past. He has published over 125 papers, and holds over 50 patents. Rajeev received his B. Tech degree from IIT Bombay, and a PhD degree in Computer Science from the University of Texas, Austin.

The 18th International Conference on Management of Data (COMAD),
14th-16th Dec, 2012 at Pune, India.