

Building Structured Web Community Portals Via Extraction, Integration, and Mass Collaboration

An-Hai Doan

University Wisconsin Madison, USA

Abstract. The World-Wide Web hosts numerous communities, each focusing on a particular topic. As such communities proliferate, so do efforts to build community portals. Most current portals are organized according to topic taxonomies. Recently, however, there has been a growing effort to build structured data portals (e.g., IMDB, Citeseer) that present a unified view of entities and relationships in the community. Such portals can prove extremely valuable in a wide range of domains. But how can we build them efficiently?

In this talk, I will present a new research vision that addresses this question. The goal is to develop a system that a small team (or ideally just one person) can quickly deploy to build an initial (but already useful) structured portal, then leverage the entire community in a mass collaboration fashion to improve and expand this portal. As such, the research agenda requires combining and extending research in information extraction, information integration, and Web 2.0 technologies, among others. This agenda is actively being pursued in the Cimple project, a joint effort between the University of Wisconsin and Yahoo Research. In the talk I will describe recent progress in Cimple, portal prototypes, lessons learned, and future directions. I will focus in particular on how Cimple raises interesting and novel challenges for both AI and database research. More information about Cimple can be found at www.cs.wisc.edu/~anhai/projects/cimple.