Craig Knoblock
University of Southern California
Participant Statement

Big ideas

Linked data: The key idea behind linked data is to publish data in a standard format and to link it to other related sources. The format used for linked data is RDF, which represents data as a set of triples consisting of subject, predicate, and object. The linking is done by assigning each entity a unique identifier and then expressing links as triples to link data across sources. Linked data makes it possible to represent data in a general way and in a larger context, which makes it easy for others to reuse the data and to understand how the data relates to other information. The idea has taken off for representing statistical data, geographic data, biological data, and cultural heritage data. We are currently working with a dozen museums to publish their data as Linked Data to create a collective of American art museums.

Big Data: While the term is overhyped and overused, there is still the core of a really important idea here, which is to exploit the tremendous computing power and disk space available today to create the infrastructure for processing massive quantities of data. This infrastructure now makes is possible to perform processing tasks that were only just recently considered to be beyond the scope of existing systems. A great example of this was Germany's use of big data infrastructure to analyze video of their team to give individuals feedback on how to improve their performance, which helped them win the World Cup in 2014. We have a project at ISI now where we are mining the data across 50 million web pages to combat human trafficking

Research Trends

Semantics of data: Driven by the recent growth of Linked Data, there is increasing work on describing the semantics of data. The semantics of a dataset are described by relating the data to an ontology for a domain. This has the advantage that data can be much more easily reused and shared. It also makes it possible to automatically integrate the data across sources and perform analysis on the data without a great deal of manual effort. In our own work, we have been developing a tool, call Karma, that automatically builds the semantic descriptions of sources using machine learning techniques.[1]

Entity Resolution/Linking: In order to combine data sources and put information in context, there is increasing work on the problem of entity resolution and linking. There has been recent work on techniques that are highly scalable and now make it possible to link data across sources in situations where there are millions of records to align. For example, we are working on a project to link mentions of names of geographic places to geonames.org, which has over 9 million records.

Data mining: There is increasing work on data mining applied to a wide variety of problems. The combination of freely available high-quality data mining tools and the wide availability of data are making it increasingly easy to exploit this technology. In our work, we are exploiting open source tools, such as Rapid Miner, to mine statistical data that is available in the Linked Data cloud.

---

[1] isi.edu/integration/karma