Dissertations

Resource-Bounded Information Acquisition and Learning

Download

Pallika H. Kanani

Date of Award
5-2012

Document Type
Open Access Dissertation

Degree Name
Doctor of Philosophy (PhD)

Degree Program
Computer Science

First Advisor
Andrew McCallum

Second Advisor
David Jensen

Third Advisor
Shlomo Zilberstein

Keywords
Information Acquisition, Information Extraction, Machine Learning, Reinforcement Learning, Resource-bounded, Web Mining

Subject Categories
Computer Sciences

Abstract
In many scenarios it is desirable to augment existing data with information acquired from an external source. For example, information from the Web can be used to fill missing values in a database or to correct errors. In many machine learning and data mining scenarios, acquiring additional feature values can lead to improved data quality and accuracy. However, there is often a cost associated with such information acquisition, and we typically need to operate under limited resources. In this thesis, I explore different aspects of Resource-bounded Information Acquisition and Learning.

The process of acquiring information from an external source involves multiple steps, such as deciding what subset of information to obtain, locating the documents that contain the required information, acquiring relevant documents, extracting the specific piece of information, and combining it with existing information to make useful decisions. The problem of Resource-bounded Information Acquisition (RBIA) involves saving resources at each stage of the information acquisition process. I explore four special cases of the RBIA problem, propose general principles for efficiently acquiring external information in real-world domains, and demonstrate their effectiveness using extensive experiments. For example, in some of these domains I show how interdependency between fields or records in the data can also be exploited to achieve cost reduction. Finally, I propose a general framework for RBIA, that takes into account the state of the database at each point of time, dynamically adapts to the results of all the steps in the acquisition process so far, as well as the properties of each step, and carries them out striving to acquire most information with least amount resources.