Aggregating Information from the Crowd and the Network

Anirban Dasgupta
Yahoo! Labs
Sunnyvale, CA
anirban@yahoo-inc.com

ABSTRACT

In social systems, information often exists in a dispersed manner, as individual opinions, local insights and preferences. In order to make a global decision however, we need to be able to aggregate such local pieces of information into a global description of the system. Such information aggregation problems are key in setting up crowdsourcing or human computation systems. How do we formally build and analyze such information aggregation systems? In this talk we will discuss three different vignettes based on the particular information aggregation problem and the "social system" that we are extracting the information from.

In our first result, we will analyze a crowdsourcing system consisting of a set of users and binary choice questions. Each user has a specific reliability that determines the user's error rate in answering the questions. We show how to give an unsupervised algorithm for aggregating the user answers in order to simultaneously derive the user expertise as well as the truth values of the questions.

Our second result will deal with the case when there is an interacting user community on a question answer forum. User preferences of quality are now expressed in terms of ("best answer" and "thumbs up/down") votes cast on each other's content. We will analyze a set of possible factors that indicate bias in user voting behavior – these factors encompass different gaming behavior, as well as other eccentricities. We address the problem of aggregating user preferences (votes) using a supervised machine learning framework to calibrate such votes. We will see that this supervised learning method of content-agnostic vote calibration can significantly improve the performance of answer ranking and expert ranking.

The last part of the talk will describe how it is possible to exploit local insights that users have about their friends in order to improve the efficiency of surveying in a (networked) population. We will describe the notion of "social sampling", where participants in a poll respond with a summary of their friends' putative responses to the poll. The analysis of social sampling leads to novel trade-off questions: the savings in the number of samples(roughly the average size of neighborhood of participants) vs. the systematic bias in the poll due to the network structure. We show bounds on the variances of few such estimators— experiments on real world networks show this to be a useful paradigm in obtaining accurate information with small number of samples.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous

Keywords

Crowdsourcing, networks, information aggregation

Short Biography

Anirban Dasgupta is a Senior Research Scientist at Yahoo! Labs. Anirban did his undergraduate studies at the Computer Science department of IIT Kharagpur, and joined Cornell CS department as a graduate student in 2000. After finishing his PhD in 2006, he joined Yahoo Research. Anirban's research interests span linear algebraic techniques for large scale data, algorithmic game theory, modeling of and algorithms for social networks and the design and analysis of randomized and approximation algorithms in general. Recently, he has also been working on information elicitation and aggregation problems for crowdsourced settings.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.