

INFORMATION AND SOCIAL SYSTEM INTERACTION

Asif-ul Haque, Ph.D.

Cornell University 2011

Ever increasing participation has made the interaction between information and social systems not only interesting to observe but essential to quantify and analyze. This dissertation presents methods for understanding such interaction through combined analysis of metadata, networks, text and log data. ArXiv, an open and highly influential scholarly communication system, served as the testbed for these methods.

In the first part of this dissertation we examine in depth interesting phenomena such as self-promotion, procrastination, visibility and geographic differences. We have confirmed the predictive power of early readership through regression and discussed undesirable effects of recommendation and possibilities of new impact metrics.

In the second part we demonstrate extraction of subtopical concepts, characterized by phrases, through a statistical method for vocabulary selection and a network based ranking. Validation via search query and click logs is advocated as relevant and scalable. A clustering scheme to summarize temporal patterns of topic clicks is also presented.

In the last part of this dissertation we present a name disambiguation algorithm and a novel evaluation method using node role based sampling in the context of network analysis. Finally we provide guidelines on performing large scale graph computation using the Map-Reduce framework.