CAREER: Computational User Modeling for Human Centric Big Data Mining

Hongning Wang (hw5x@virginia.edu)
University of Virginia

Discovering actionable knowledge from big data has generated enormous impact to our society: e.g, improved health care, enhanced homeland security, and accelerated scientific discovery. It is worth noting that a large portion of data is produced and consumed by humans, and it will continue to fuel exponential growth. Comparing to the non-human-generated data, e.g., sensor data and bio-sequence data, which is usually associated with predefined schemes and structures, human-generated data is mostly unstructured and thus not directly accessible by machines, posing significant challenges for data mining. Yet, human-generated data contains rich semantic information and reflects the latent intentions of people, thus they are remarkably valuable for knowledge discovery. In order to effectively extract actionable knowledge from such human-generated data and facilitate end-users to exploit the mined knowledge, we need to put human into the loop of mining big data via proper computational user modeling.

In this project, we propose to tackle the research problem of "human-centric big data mining" from three different perspectives. First, ***User Understanding***. To effectively discover actionable knowledge from human-generated data and optimize users' decision making process by the mined knowledge, we propose to fuse solutions from both computational methodologies and behavioural science and psychology studies. In particular, we plan to build a general computational framework based on probabilistic generative models for modeling human behaviors. The dependency structure and prior in this proposed generative model will be specified by the established theories from the fields of behavioural science and psychology studies. Second, ***Knowledge Representation***. Analyzing human-generated data in an ad-hoc manner is far from ideal, which will inevitably isolate the mined knowledge and hamper its usage across different applications. We plan to develop a general knowledge representation scheme, e.g., an ontology structure, based on the theories from behavioural and psychology studies to organize the mined knowledge about human behaviors, so that the learned knowledge can be used in multiple application scenarios transparently. Third, ***Intelligent Knowledge Service Techniques***. Game-theoretic models will be explored for building an algorithmic framework that actively interacts with the users and learns from them, e.g., modeling the interaction process as a multi-agent game. This enables the system and users collaborate for knowledge acquisition and decision making.

The **intellectual merit** of this project consists of multiple aspects. First, we propose a general computational framework for modeling human-generated data. The generation process of the data will be explicitly modeled, and it helps us discover the underlying structure of human behaviors, e.g., users' latent intent when generating the data. Second, the mined knowledge will be directly fed back to the knowledge service system for optimizing users' decision making process. Due to the in-depth understanding of users' intention gained from the proposed user modeling approach, we can better understand users' information need and decision preferences, and so as to maximize the utility of the mined knowledge. Third, the proposed research aims at developing solid computational models for user modeling based on the theories and studies in psychology, behaviorial science, and cognitive science. Possibilities of enhancing quantitative research with qualitative studies will be explored on multiple application scenarios in this project.

The **broad impact** of this project follows from both improving the utility of mined knowledge from big data, and catalyzing the integrated research across multiple disciplines. Given the humans are the producer and consumer of big data, the depth of user understanding directly determines the upper bound of optimality that any automated mining algorithms can achieve in mining big data. The research proposed in this project will raise this upper bound via explicit modeling the data generation process and identifying the latent users' intents behind their generated data. More importantly, such a study will never be just an one-sided effort; the developed method and mined knowledge about humans will also benefit the studies in psychology, cognitive science, design, and more. At the educational level, the proposed research will be adopted into course materials and projects for graduate and undergraduate students, and K-12 outreach activities.

Keywords: user modeling, data mining, big data, generative models, knowledge representation, game-theoretic models, and interactive behavior modeling.