Thesis Proposal

Institute for Software Research
Societal Computing

Learning User Latent Attributes on Social Media

Binxuan Huang

Thursday, August 22, 2019
1:00 PM - 4:00 PM
Wean Hall 4220

In recent years, there is a growing interest in using social media to understand social phenomena. Researchers have demonstrated many important applications of user activity understanding in online social media, such as presidential election prediction, earthquake early detection, and disaster management. A social media site is mixed with users with various attributes, in terms of gender, location, political affiliation, social roles, and etc. Different types of users may have different motivations, different opinions towards certain topics, different resources at their disposal, different behaviors in events. If researchers want to understand what is happening on a social media site, it is important to know where a post comes from, who wrote this post, and which party the author belongs to.

In this thesis, the goal is to predict users' latent attributes such as their locations, social identities, and political orientations. Many times, these attributes are not explicitly provided by users. For example, there are no ways for users to specify their ideologies on Twitter. Even though there is a location field in a Twitter user's profile, it is often empty or filled with noisy information unrelated with any geographical location. However, many times online users would unavoidably provide indicative clues which are helpful for identifying their attributes in their posts. For example, from a post "food in cmu is delicious :)", we can infer that this user is probably living in Pittsburgh.

Thanks to the massive text data on social media, we can learn rich knowledge from text to predict users' attributes. In the meanwhile, text data from social media often comes with a significant amount of meta data. Take Twitter for example, a single tweet object contains one short text with multiple meta fields like posting time, tweet language, user's personal description, and etc. How to efficiently handle text data combined with meta information still needs to be considered. Furthermore, data from social networks also contains rich connection information, eg. mentioning, following. It is still a challenge task to combine text, meta data, user network together for user attributes prediction.

In this thesis, I will approach user attributes prediction at three levels --- single post, user timeline, graph-level classification. First, I will present a global location prediction system that uses one single tweet as input to learn one user's location. It utilizes location-related features in a tweet, such as text and user profile metadata. Second, I plan to extend the tweet-level prediction system to user-level, which combines multiple posts in one user's timeline. I will demonstrate the effectiveness of this model on the task of user social identity classification. An improved user-level hierarchical location prediction model will also be presented. In these described models, I mainly focus on learning user attributes from users themselves. In the next step, I will consider social graph as additional information to improve performance. Users connected in a social network often show similarities in certain aspects, which is a well-known phenomenon called social homophily. Last, to reduce the computation cost for learning each individual attribute separately, I intend to propose a multi-task learning system that learns all attributes jointly.

**Committee: Dr. Kathleen Carley (Chair), Dr. Yulia Tsvetkov,
Dr. Zico Kolter, Dr. Huan Liu (Arizona State University)**

isr institute for SOFTWARE RESEARCH | Carnegie Mellon University School of Computer Science