Research Statement

Pavan Kapanipathi, Kno.e.sis Center, Wright State University, USA (pavan@knoesis.org)

Social media has become a part of our everyday life - just of them attracts one billion users every day. This extensive usage and participation enabled by these platforms have attracted researchers in every domain to tap into the wisdom of the crowd. For examples, tasks such as understanding users to customize their content on the web, disaster management, and analyzing social issues extensively harness social data. Furthermore, for users on the internet, these platforms are becoming increasingly common for information seeking and consumption. However, as part of the growing popularity, information overload poses a significant challenge. While social media is being utilized and analyzed by a broad spectrum of researchers, there is yet remains a tremendous scope of research to improve the analysis for each domain of interest.

Current Research.

My research interests encompasses the areas of Semantic-empowered Intelligent Systems (Semantic Web, Knowledge Graphs), Social Data Filtering/Analysis/Mining, and User Modeling for Personalization and Recommendation. Specifically my dissertation addresses the information overload challenge on social media. I use the knowledge available on the web with statistical and Information Retrieval techniques to address the following research challenges:

1. Due to its short-text and informal nature, social media lacks semantic context for analyzing user behavior and interests. How can we harness openly available knowledge bases to augment social media posts with semantic context? [Modeling – C1, P1]
2. While the location of a user plays a prominent role in many applications such as personalization (location-specific personalization and recommendation) and crisis response (location-based information diffusion), less than 4% of Twitter users share their location on Twitter. Traditional approaches use supervised learning techniques to estimate user location which are (1) time-intensive and (2) data-driven, hence not quickly adaptable to new locations. How can we substitute training data with domain-specific knowledge bases to build an unsupervised approach for location prediction of social media users? [*User Modeling – C2*]
3. Social data is filtered using keywords as queries. However, streaming data with keywords as filters is insufficient to gain access to a relevant, large pool of information. How can we expand the queries using background knowledge to improve the recall of relevant tweets? [*Query Expansion, Real-time Information Filtering – C3, C4, D1*]
4. Tracking streaming social data relevant to dynamically evolving topics is challenging, specifically due to the evolving vocabulary representing the topics. For instance, a disaster like hurricane sandy spans different locations at different points in time. This

is reflected with the shift in vocabulary used on social media. Therefore, monitoring such topics requires a setting that adapts the filters with the evolution of the topic. I focused on accomplishing this goal by utilizing an timely, updated crowdsourced knowledge base for dynamically updating the topic-filters for social data. [*Query Expansion, Real-time Information Filtering – T1*]

Impact and Applications. Twarql, a system which is an outcome of our work addressing research challenge 3, is open-sourced and has also won the Triplification Challenge in open track. Furthermore, the annotation pipeline of Twarql has been incorporated into Twitris for in-depth analysis. Work on the location prediction of social media users and dynamically tracking events has significantly contributed to various grant proposals to all the top agencies, including the NSF, NIH, and DARPA, in addition to providing proposal writing experience as a participant for 3 award winning proposals at Kno.e.sis (2 NSF and 1 NIH). The approach developed to address research challenge 1 involved building a hierarchy of concepts from a crowdsourced knowledge base. More than 5 research labs worldwide are utilizing the hierarchy built for different purposes.

Future Research Plan.

My long-term research plan is to explore the use of knowledge bases of diverse domain for personal, collaborative, and social data that people across the world publish on the Web. I will continue to pursue research on addressing challenges with the filtering and retrieval of such data and applying my research to diverse domains. This in turn would provide opportunities for exploring and collaborating with researchers from other domains such as healthcare and disaster management. I plan to contribute to the areas of Data Mining, Semantic Web, Intelligent Systems, and Social Data Analysis, pursuing interdisciplinary collaborations whenever possible for use cases and data, applications and real-world impact.

Knowledge-aware social data collection and analysis. Social computing has grown in popularity in diverse domains to gain insights into the online world. Presently, at Kno.e.sis we are utilizing social data for at the least three domains: (1) Epidemiology: To analyze and understand the use of marijuana across United States (). (2) Psychology and Sociology: To determine and evaluate potential harassment and harassers and to determine and predict users with depression (). (3) Disaster management: To analyze, understand, and predict the extent of damage to infrastructures during natural disasters (). While contributing to some of these projects and proposals, the analysis requires significant understanding of the domain, and that is provided by the domain-experts. This follows a quote by Lenat and Feigenbaum, ”If a program is to perform a complex task well, it must know a great deal about the world in which it operates.” With this motivation, I am inclined to research and develop a framework that can help create and utilize domain-specific knowledge for social computing. This will require the exploration of Semantic Web (knowledge graphs and ontologies) based methods to represent common concepts in ontologies along with the help of domain experts, use of information extraction techniques to transform unstructured social data into structured forms, and the exploration of weighted aggregation methods for the

integration of social data and data from other sources.

User-specific Information Extraction from Online Textual Data: Customizing the web, either via personalization or recommendation necessitates understanding user interests. While most recommendation systems look through explicit ratings provided by users (stars on Netflix, and Amazon), the textual data generated by users provide a larger ground for analyzing user needs and interests. While, it is important to acknowledge that there has been research done in this area, there still exists enough research problems to be addressed on new forms of data such as social data. One of my short-term goals is to extract the mundane activities users are interested in from their Twitter data. For example, user interests that are hobbies such as running, sky diving, and cooking can be modeled and utilized for better recommendation.

Sensor and Textual Data integration and analysis for Intelligent Systems. While social and sensor data are two of the prominent real-time data streams available, both are largely analyzed separately to solve real-world problem – be it in the domain of traffic analytics, healthcare, and disaster management. It is necessary to design multi-dimensional, cross-modal aggregation and inference techniques to compensate, complement, and corroborate analysis of each of the data modalities distinctly. For example, reasoning a traffic jam (concluded from sensor data) using social data, or compensating fallacious weather sensor data during disasters using variation of domain relevant terms on social data.