RESEARCH SUMMARY

MEENA NAGARAJAN

Problems Investigated, Datasets Used, Relevant Publications

Over the last few years, there has been a growing public fascination with 'social media' and its role in modern society. At the heart of this fascination is the ability for users to create and share content via a variety of platforms such as blogs, micro-blogs, collaborative wikis, multimedia sharing sites, social networking sites etc. The volume and variety of user-generated content (UGC) and the user participation network behind it are creating new opportunities for understanding web-based practices and building socially intelligent and personalized applications. Investigations around social data can be broadly categorized along the following dimensions: (a) understanding aspects of the user-generated *content* (b) modeling and observing the user *network* that the content is generated in and (c) characterizing *individuals* and *groups* that produce and consume the content. My interest in this space is driven by the need to understand combined effects of the content, people and network dimensions on various emerging social phenomena on the Web. My dissertation research primarily focuses on the analysis of various aspects of user-generated content that are central to understanding inter-personal communication on social media. The objective of my work is to bring structure and organization to unstructured chatter on social media centered around the following questions:

- ***What* are people talking about:** What are the Named Entities and topics that people are making references to? How are cultures interpreting any situation in local contexts and supporting them in their variable observations on a social medium?
- ***How* are they expressing themselves:** What do word usages tell us about an active population or about individual allegiances or non-conformity to group practices?
- ***Why* do they scribe:** What are the diverse intentions that produce the diverse content on social media? Can we understand why we share by looking at what we predominantly do with the medium? What emotions are people sharing about something?

WHAT do people write

Cultural Named Entity Extraction [ISWC09a] [AAAI2010], Summaries of Social Perceptions [WISE09, ISWC09b]

While there is a rich body of previous work in processing textual content, certain characteristics of UGC on social media introduce challenges in their analyses. A large portion of language found in UGC is in the Informal English domain — a blend of abbreviations, slang and context specific terms; lacking in sufficient context and regularities and delivered with an indifferent approach to grammar and spelling. Traditional content analysis techniques developed for a more formal genre like news, Wikipedia or scientific articles do not translate effectively to UGC. Consequently, well-understood problems such as information extraction, search or monetization on the Web are facing pertinent challenges owing to this new class of textual data.

HOW do people write

Self-Presentation in Online Dating Profiles [ICWSM09]

My dissertation research tackles some of these challenges in the analysis of textual UGC, by supplementing statistical Natural Language Processing (NLP) techniques and Machine Learning (ML) algorithms with *contextual information* from the social medium and external resources like domain Ontologies, Taxonomies and Dictionaries. In my work, I have examined a variety of research problems using data from popular social media platforms (left inset). Much of my work has greatly benefited from collaborations with researchers at industry (HP Labs, IBM Research - Almaden, Microsoft Research) and academic labs (Kno.e.sis Center and Dept. of Cognitive Psychology at Wright State, Information School at UC-Berkeley). My work was also the basis of two competitive grants from IBM and Microsoft[1]. Some results of my work have been absorbed into two deployed social intelligence Web applications[2] - BBC SoundIndex, that measures the pulse of a populace using UGC from online music communities; and Twitris, that aggregates user perceptions behind real-time events using data from Twitter.

WHY do people write

Mining Monetizable User Intentions [WI09], Mining User Opinions [VLDB09]

While my dissertation focus has been in the analysis of UGC, I see an exciting opportunity to study the synergy between the micro-level variables of UGC (what, why, how), the network structure (who is connected to whom) and people characteristics (poster characteristics), in understanding complex social phenomena on the Web. For example, how does the interplay of the *topic* of discussion, *emotional charge* of a conversation, the presence of an *expert* and *connectedness* between participants, together affect emerging social order in an online conversation, explain information propagation in a social network etc.?

---

[1] IBM UIMA Innovation Award 2007, Microsoft's Beyond Search – Semantic Computing and Internet Economics Awards, 2008

[2] BBC SoundIndex - a catologue of popular music artists and tracks generated from online music communities. Joint work with IBM Almaden. http://www.almaden.ibm.com/cs/projects/iis/sound/; Twitris – Summarizing Twitter across space, theme and time dimensions. Joint work with researchers at Kno.e.sis. http://twitris.knoesis.org