Abstract

The emergence of social discussion platforms has introduced a disruptive change in how information and discussion topics reach people. Interactive online networks such as Facebook, Youtube, Pinterest etc., and microblog platforms, such as Twitter, have been widely adopted by end-users of social media. It is immensely valuable to analyze the user-generated content, the underlying user network, and user attributes, to form a better understanding of the social information movement dynamics, which is the overall aim of this work.

The thesis performs analysis of Twitter text (tweets) and social connections (followships), from the perspective of modeling correlations between (a) social connections and structures such as communities, (b) the lifecycle of topics of discussion, and (c) the movement of topical information on the social networks. In this work, we provide a novel definition of topics using user-given explicit hashtags, latent semantic content and temporal continuity of the content, and empirically discover and model several interesting and correlated attributes of such topics. This includes analyzing the lifecycle of these topics, assessing user roles in the lifecycle, finding the impact of social communities in such lifecycles, finding topical homophily and discovering its characteristics, modeling topical information flow on the social network and assigning hashtags to tweets for better topic discovery. Prior to that, we establish in the background that a correlation exists between the language usage of individual users of social networks, and the social communities they implicitly belong to. In one direction, we observe that, language used by individuals is influenced by the language used by the social communities they belong to. In the other direction, we empirically show that, language usage similarity is a reasonable approximator for social communities. Subsequently, the core of the work is laid out over the following research modules.

Topic Finding and Topic Lifecycle Analysis: Since topics are not well-defined in Twitter and there is no universally accepted external definition, we propose a novel definition of topics, using a combination of user-given explicit hashtags, implicit (latent) content semantics and temporal continuity of content semantics. Further, we model the lifecycle of topics thus defined, and perform an exploration towards user participation in the topics. The lifecycle analysis addresses how these topics get born, become prominent, and evolve in terms of its constituents (thereby, morph), rather than simply getting born and dying too soon. The user participation angle of our analysis shows how the behavior characteristics of the participants play roles in shaping up the lifecycle of topics thus derived.

Topical Homophily and Its Characterization: Another core contribution made by the present work is towards defining and identifying topical homophily, and characterizing such homophily. As per the classical definition of homophily, “similarity breeds familiarity”. Our work defines topical homophily as familiarity being bred by user pair similarity computed over their participation in similar topics. We empirically demonstrate the existence of topical homophily, and further show that such homophily is characterized by it non-monotonic nature, as well as, it peaks faster for tighter definitions of topical similarity of semantic concepts.

Topic Lifecycle as an Outcome of Social Community Dynamics: The novel definition of topics, and the significant presence of topical homophily at different levels of familiarity, motivate us to investigate whether social communities, rather than only individual users in isolation (discounting social connections), have a role to play in lifecycle of topics. We show that, topics associate with communities at any given time, by modeling the correlation between social communities and the constituent hashtags of the topics. We show that the constituent hashtags are independently used across communities, and the topics evolve atomically (independently) within communities. Empirical
demonstration of the lifecycle of topics as a property of social community dynamics is novel in the literature.

**Topical Information Diffusion:** Information movement is one of the core properties of social networks. The prior works mainly account for two factors: (a) the movement of information along the social connections and (b) individual users getting sufficiently excited (“activated”) to move the received information further forward. On the other hand, we believe that whether or not a user chooses to forward the information they receive will also depend upon the user’s affinity towards the tweet information topic content. We thereby propose **TopSPA**, a novel information diffusion model that underneath uses the spreading activation (SPA) technique, but blends it with user affinity towards its information content.

**Topic Augmentation:** While the use of hashtags is integral to our definition of topics, a large fraction of the tweets does not contain any hashtag. This requires assigning hashtags to such tweets, which is equivalent to recommending hashtags for tweets. We propose a model using a combination of three components: NLP, a temporal factor and a socio-temporal factor. The NLP factor analyzes the semantic content of the test tweet with respect to the concept space covered known training tweets. The socio-temporal factor assumes that, individuals are more likely to use hashtags that have been used more recently, by their stronger social connections. The temporal factor assumes that, new topics that break out as bursts in real life are likely to reach individual users from other sources, such as the outside world, instead of via social friends.

Our work gives its end-user to an end-to-end execution pipeline to correlate social connections, topics and information diffusion on Twitter. The key contribution of this thesis is an empirical establishment of two fundamental hypotheses: (a) topics and their lifecycles are community-level properties, and (b) information diffusion is governed by the topic of the content.