Definition and Multi-Dimensional Comparative Analysis of Ad Hoc Communities in Twitter

Sofus A. Macskassy
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292
sofmac@isi.edu

Abstract

We here present an early-stage prototype tool for defining and analyzing communities in Twitter. The tool takes a set of Twitter users and profiles them based on their tweets. This profiling is based on earlier work, where we map entities mentioned in tweets to Wikipedia entries, which in turn lets us profile a user based on the Wikipedia categories are related to his or her tweets. From here, we can define ad hoc topic-based communities (e.g., all users who discuss Wikipedia topic K). The tool is focused on contrast analysis, where we have baseline behavior or another community to compare against.

Motivation

Twitter and other social media are extremely rich and dynamic in terms of content and online social behaviors. Much of what goes on in these online platforms reflect events in the real world and the social dynamics and social networks are often treated as partial observations of the networks in the real world. If we could monitor and analyze these online streams then we could get a real-time assessment or "pulse" of the world.

As a consequence, the area of social media applications is vibrant and chaotic with new tools being created every day to help users and analysts in various forms make sense of what is going on. Many of these tools focus on high-level dynamics such as tracking the volume or "tag cloud" around particular events or other broad-level dynamics.

This prototype demo is slightly different in that we are focusing on digging deeper into the analysis of a particular group or community. As such, the demo is focuses on two primary tasks: (1) defining a group (or community) and (2) monitoring and analyzing this group over time. The demo focuses specifically on *contrasting* analysis, with the goal of comparing two different communities.

Task 1: *Ad hoc* definition of a community

Our approach to defining a community is relatively loose. Specifically, we realize that the notion of a community is very context dependent and so we wanted to make it easier to define a community based on certain characteristics of the people one might want to include, whether they be demographic or psychographic. For example, one might be interested in all soccer moms in the Chicago suburbs, or all politically active people in Islamabad. To this end, we have developed technology to profile people based on the content they generate (Michelson and Macskassy, 2010). We do this by first identifying entities users mention in their posts and then mapping them into an ontology such as Wikipedia. This will result in a "tree" of ontology nodes. We generate what we call a psychographic profile by aggregating all these trees. This profile generation process is shown in Figure 1. Once profiles have been generated, we can define a community as a query over these profiles (e.g., all users who talk significantly about soccer).

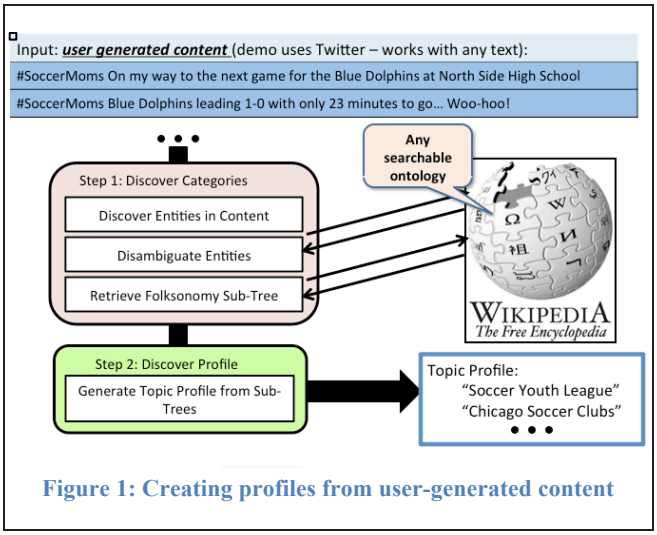

Figure 1: Creating profiles from user-generated content

Task 2: Contrasting analysis of communities

Once communities have been defined, we first show a list of the users that make up each community; ranked by how much content they have produced (see Figure 2).

From here, the user can then get a comparative analysis of the two communities along a set of dimensions.

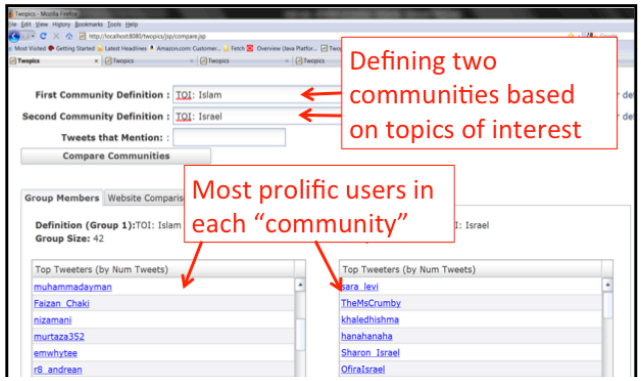

Figure 2: Defining two communities of interest

Identifying web-sites of interest

In this analysis, we look at the websites the users from each community link to and then identify three categories of websites. We first de-reference all shortened URLs to identify the real website, then compute statistics on linking behavior from each community, and finally categorize a website into one of three categories:

1) **Unique**: only one community links to that web-site.
2) **Significant**: one community links to the web-site significantly more often than the other.
3) **Neither**: there is no statistically significant difference in linking behavior.

Figure 3: Identifying web-sites of interest

If both communities link to a site, then we use a *Z-test* to compute statistical significance. This is based on the size of the communities and the number of users linking to the website. Figure 3 shows a screen shot from the demo.

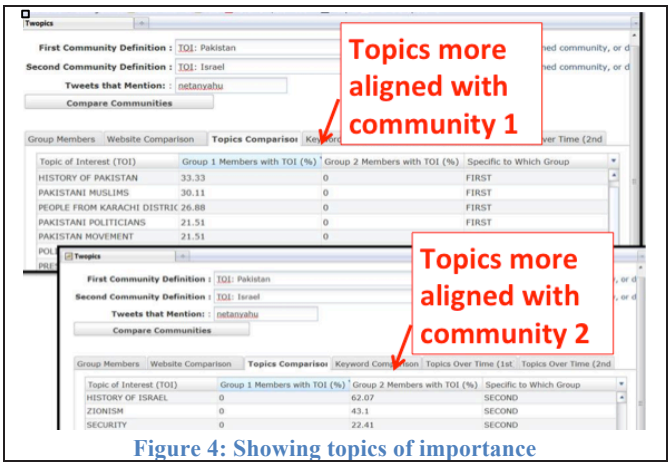

Figure 4: Showing topics of importance

Identifying topics of importance for each community

The next dimension of analysis is at the topic-level. We analyze the topical concepts in each tweet in the given community (where a topic is an ontological node), and compute whether it is statistically more aligned with one or the other community in a similar manner as we did with the website comparison just described (again using a *Z-test*). A screen shot is shown in Figure 4.

Ad Hoc Chatter Analysis

Often one would like to know what a community is saying about a topic or entity. Our demo shows how one can query on a keyword or phrase and it will on the fly compute words which each community use in the context of that phrase. We apply machine learning to learn to categorize tweets and then extract the features (words) the model uses to discriminate between the two. We then cluster features and tweets to give users an overview of their differences. (screen shot omitted due to space).

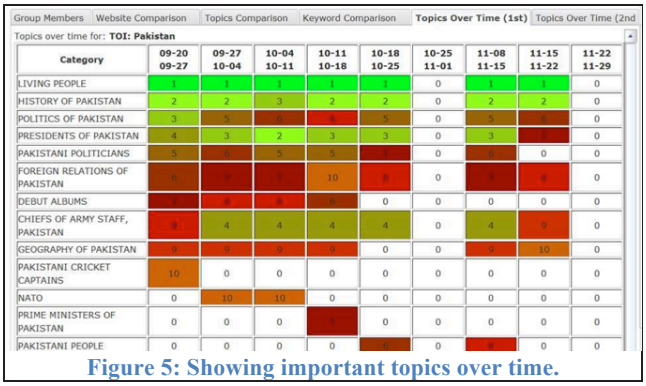

Figure 5: Showing important topics over time.

References

Michelson, M and Macskassy, S. A. (2010). Discovering Users' Topics of Interest on Twitter: A First Look. *Proceedings of the Workshop on Analytics for Noisy, Unstructured Text Data (AND)*.