i290 - Social Computing. Group 9, Sebastian Benthall and Sean Chen

Problem Definition

Our project will involve a series of experiments to test whether and how social network structure is related to the content contained in its nodes. In particular, we will focus on Twitter, and attempt to find regularities in the relationship between the network structure of 'followers' with the subject matter or 'topics' of tweets.

Research questions include:

- Is there any correlation (or anticorrelation) between diversity of topics and number of followers? In number of people followed?
- Do clusters form around certain topics? Do these clusters intersect in any regular ways?
- (Advanced question, time permitting) Is there any regular relationship between the social distribution of topics across the network and the semantic hierarchy of topics (as elaborated by, e.g., WordNet)
- (Advanced question) Is there regularity to the way topics propogate across the social network over time?

There are many directions this work could take depending on the results of preliminary steps. In our project, we will approach them time-permitting and as our methodology evolves.

For this research, we will apply probabilistic topic modeling[1] methods to extract topics from twitter text.

Motivation

Aside from the intrinsic academic interest in arriving at a better understand of the structure of social networks, the answers to these questions could have industrial value as part of recommendation systems, advertising targeting, or social organizing tools.

This research addresses 'social' questions by taking a tool used for document analysis and applying it to documents (tweets) written and consumed in an explicitly social way.

Approach, Algorithm, Data

Our first pass at a methodology will involve:

1. Get a sample subset of the Twitter social graph.
2. Get the log of tweets for each user in that social graph.

---

[1]See Steyvers, M. & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), *Latent Semantic Analysis: A Road to Meaning.* Laurence Erlbaum

3. On the total data set of tweets, run LDA to generate a topic model of the total set. Use this to determine the total distribution of topics across the graph. For our first pass, we will use the MALLET software package to perform the topic modelling.[2]
4. Apply that topic model to each individual *user's* tweets to determine their individual distribution of topics.
5. Analyze this derived network data. For example, to test the effect of topic diversity on number of followers, we would check for correlations between the information theoretic entropy of a node's topic distribution and its in-degree.
6. Repeat for multiple sample Twitter subgraphs to test validity of hypotheses.

Based on an promising leads from this initial analysis, we will work to refine our method to address issues of unbiased sampling, computational tractability, data cleaning, and the feature representation of tweets.

As we develop experimental hypotheses through the initial investigation, we will verify them by applying our methods to broader data sets. Since Twitter data is so plentiful, we will have no problem keeping "testing" data separate from "training" data.

Moving on this initial hypothesis, we will use the methods we've developed in the first phase to study the distribution of topics across sample network structures. We will approach the plausible hypothesis that clusters of twitter users are to some extent characterized by the topics they tweet about. For example, entrepreneurs will follow other entrepreneurs who will all tweet about topics related to entrepreneurship, while political buffs will follow other political buffs that tweet about political topics.

There are three approaches to this problem we could employ. The first would be to develop a way to visualize the distribution of topics across a social graph by, for example, mapping from topics to the color spectrum and coloring nodes based on their topic distribution. Such visualizations could help us get an intuitive sense of the validity of our hypothesis.

Time permitting, if this visualization proves promising, we could attempt to verify the intuitive result statistically. (For this we would need to do further research into the available techniques.)

Lastly, if there are significant regularities to graph structure and topic distribution, we may try to develop a generative model of contentful, 'twitter-like' graphs, based on the LDA probability distribution. Ideally, this model could be compared favorably in relevant dimensions to real Twitter data.

Milestones

Project milestones may include:

1. Proof of concept: evaluation of correlation between topic entropy and in-degree of twitter feeds, based on small data sample.
2. Verification of evaluation based on larger data sample.
3. Methodological refinement, possibly including natural language processing or performance optimizations.

---

[2]See http://mallet.cs.umass.edu/topics.php

4. Visualization tool for showing distribution of topics over social network, for intuitive view of topic-based clustering. Form clustering hypothesis.
5. Statistical analysis of topic-based clustering hypothesis.
6. Develop generative model of topic-based network structure.
7. Report.