11719 Assignment 3

Seungwhan (Shane) Moon, Saloni Potdar, Lara Martin

March 17, 2014

1 Introduction

Studying influence in social networks is an important topic that has attracted the attention of a variety of researchers in different domains. People often seek the opinion and advice of their peers regarding various decisions. This behavior gives rise to a certain set of individuals in the social network, referred to as influentials or leaders, who have a huge impact on other peoples opinions, actions and behavior [1]. In the context of this project, we define leaders as important individuals who remarkably influence the language usage and its propagation process. We wish to identify the leaders as role models who inspire imitation especially through language. Our leader identification is similar to authority identification in CQA(Community Question Answering). It is the task of identifying users who can provide a large number of high quality, complete, and reliable answers [2]. There have been various approaches that have been used to identify opinion leaders in the a community forum. They use PageRank, HITS, InDegree and ExpertiseRank or other non-linguistic feature like Longevity (how long the person has stayed on the forum), Activity(number on posts by the user), Influence (using the set of repliers for document and the number of replying users for document), Novelty, the interest and expertise in a topic by looking at documents a user has posted or commented on a specific topic looking at the representative terms on a specified topic a user has employed to determine the extent of expertise [3]. In this paper, we propose several approaches to identify leaders solely based on the linguistic features of the text.

2 Proposed Methods

2.1 Influence Propagation through Language Accommodation

In this section, we propose a method to identify student leaders by analyzing the propagation of influence through language accommodation. Given a series of threads on an online discussion forum, we are interested in studying how a student leader coordinates other members' language usage, and how the language accommodation is propagated to the subsequent members via both direct and indirect interaction with the student leader. Suppose, for instance, that during an online discussion a person b coordinates towards a with respect to a specific linguistic style marker m, and that within a short period of time, we find an evidence that another person c coordinates towards b on the same marker m. We argue that c should be considered as pertaining to the influence graph of a, contributing to the evidence that a is a leader.

2.1.1 Related Work

The task of discovering leaders of a community has been well studied, especially in the domain of social networks. [4, 3], for example, provides a great framework for measuring influence propagation on a social network using a frequent pattern mining approach. The assumption made in this work is that certain community actions (e.g. sharing an article) performed by a user are visible to the members of the network, which then influences these members to perform the same actions. The proposed method analyzes the

timestamps at which each action has occurred, and determines which users are leaders when it comes to setting the trend for performing certain actions.

We extend this framework to our study of student leaders. Specifically, we consider language accommodation as a primary action of users, which would allow us to focus on the domain-independent features of a student leader. [5] proposes a concise probabilistic coordination measure which defines language coordination from a speaker to a target on a specific word. We formulate our language influence based off the coordination measure proposed in [5].

2.1.2 Problem Definition

We first define an undirected interaction graph $G_\tau = (V, E)$, where the nodes are the authors, and the edges between authors u_i and u_j represent the degree of interaction between the authors during the observation period τ. For example, if users u_i and u_j have participated in the same threads of discussion frequently, the edge that connects them should be strong. We consider different observation periods τ to study the time-varying aspect of the influence graphs over the course of the community. We borrow the definition of an *action log* from [4], which is a relation *Actions(User, Action, Time)* that contains a tuple (u, a, t) indicating that author u performed action a at time t. In our study, each action of an author refers to language accommodation acted by its neighbors during their interactions (e.g. an exchange of conversation). Specifically, we define $c^m_{j\to i} \in \mathbb{R}$ be a measure of an action $a^m_{j\to i}$ that a speaker u_j coordinates towards a target u_i with respect to a linguistic style marker m within a pair of utterances. We use the similar coordination measure proposed in [5] for estimating $c^m_{j\to i}$, where the probabilities are estimated over utterances during the period τ.

We then define an action propagation as follows: we say that an action $a^m \in A$ propagates from author v_i to v_j *iff* $(v_i, v_j) \in E$ and $\exists (v_i, a^m_{j\to i}, t_i), (v_j, a^m_{k\to j}, t_j) \in Actions$ with $t_i < t_j$, $c^m > \gamma$ for an accommodation measure threshold γ, and for some k.

For each action a, we construct a propagation graph $PG(a) = ((V(a), E(a))$, where $V(a) = \{v|(v, a, t) \in Actions\}$, and $E(a) = \{t_j - t_i > 0|(v_i, a, t_i), (v_j, a, t_j) \in Actions\}$. Given the propagation graph PG, we define the user influence graph $Inf_\pi(u, a)$ as the subgraph of $PG(a)$ rooted at u, of which the nodes are reachable within the propagation time threshold π. Therefore, the task of discovering a leader can be framed as finding an $Inf_\pi(u, a)$ with the maximum size.

2.2 Topic Models

We propose two methods to tackle the problem of identifying leaders using topic models.

Our assumption while considering this problem is that the student leaders talk like instructors. They have more knowledge as compared to other students. The language used by student leaders is similar to those of instructors - i.e. they explain the concept, are more authoritative, and try to help the other students understand the concept better. Thus our assumption is valid as instructors and students will most likely use similar words to fulfill their overlapping goal. They are also less likely to be off-topic like the other students, since they are aware of the functioning of the forum.

Therefore, the classification problem of whether a student is a leader or not can be framed as measuring the likelihood of an utterance of the student generated from the topic model trained on the utterances by instructors. We can then formulate the problem as a text categorization task using topic models:

$$\hat{M} = \underset{M \in \{M_I \cup M_S\}}{\operatorname{argmax}} \sum_{u \in U} \mathcal{L}(u|M) \tag{1}$$

where M_I and M_S are the topic models trained on the utterances by the instructors and the students, respectively. U_a is a set of utterances by a particular student a. The approach of using topic models in text categorization is known to be effective [6].

The language of a person changes over time as he learns a particular subject. This makes the influence modeling difficult with just topic modelling over all the different participants. Thus the consideration of time is an important factor while approaching this method of identifying student leaders. To address this

we propose a time varying topic modeling using LDA. Our basis for identifying leaders is that the topic distribution of student leaders and the instructors will be similar. A LDA-model is built for the instructors and TAs and for each student individually. We then use a similarity measure like KL divergence to find out how similar the topic distribution is for a student and the instructors which will help us to find out whether the students are leaders or not.

2.3 Syntactic and Lexical Features

In our final analysis, we plan to identify student leaders by looking at the posts through a much smaller lens. We will be looking at the words themselves. We believe that instructors use different word choices than TAs, who use different words than students. We also believe that people within a group will use words similarly. In essence, the linguistic style of the different groups will vary.

Since the student leaders generally have something to offer to people (i.e. knowledge), we suspect that the other students will tend to be more polite toward the student leaders, and using indirect forms of questions, in order to extract the information [7]. By use of the LIWC[1] and General Inquirer[2] word corpora, we will extract the parts of speech, emotions, and various, broad semantic categories from the words. Similar methods, including the use of the LIWC corpus, have been used previously in research that predicted the level of power of a speaker [5]. These categories will also help us to investigate politeness.

Overall, there are two methods we are looking at to approach this problem, using these data sets. One way is to extract the labels for the instructors, TAs, and students, hand-labeling student leaders, and then training different types of classifiers in order to create a model that can most accurately predict which category the author of the post is likely to belong to. Similar work has had reasonable success using SVMs [7]. Another way is to cluster posts with similar styles together. These clusters will be created based on the number of categories that the words being compared share. Then we will see if the clusters correspond to instructors/TAs/student leaders/other students, or however many clusters is deemed the most suitable. It is unclear at the moment which method would be more valuable for our research.

3 Evaluation Plan

We plan to evaluate each of the proposed methods on the prediction task of student leaders. Because the labels for student leaders are not available, we instead plan on presenting other indications of student leaders such as the correlation between the number of votes and the leadership. The correlations within the three proposed methods will also be presented.

References

[1] H. Sharara, L. Getoor, and M. Norton, "Active surveying: A probabilistic approach for identifying key opinion leaders," in *The 22nd International Joint Conference on Artificial Intelligence (IJCAI '11)*, 2011.

[2] A. Pal and J. A. Konstan, "Expert identification in community question answering: Exploring question selection bias," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, (New York, NY, USA), pp. 1505–1508, ACM, 2010.

[3] Y. Li, S. Ma, Y. Zhang, R. Huang, and Kinshuk, "An improved mix framework for opinion leader identification in online learning communities," *Knowledge-Based Systems*, vol. 43, no. 0, pp. 43 – 51, 2013.

[4] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Discovering leaders from community actions," *CIKM '08*, 2008.

[1] http://www.liwc.net/

[2] http://www.wjh.harvard.edu/~inquirer/

[5] C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg, "Echoes of power: Language effects and power differences in social interaction," *Proceedings of WWW 2012*, 2012.

[6] M. Chen, X. Jin, and D. Shen, "Short text classication improved by learning multi-granularity topics," *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[7] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts, "A computational approach to politeness with application to social factors," *CoRR*, vol. abs/1306.6078, 2013.