11719 Assignment 4

Seungwhan (Shane) Moon, Saloni Potdar, Lara Martin

March 26, 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 $\exists (v_i, a^m_{j\to i}, t_i), (v_j, a^m_{k\to j}, t_j) \in \mathit{Actions}$ 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 $V(a) = \{v|(v, a, t) \in \mathit{Actions}\}$, and $E(a) = \{t_j - t_i > 0|(v_i, a, t_i), (v_j, a, t_j) \in \mathit{Actions}\}$. 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.

2.1.3 Preliminary Analysis

In the formulation of the proposed method, we have made two main hypotheses as follows:

1. Non-leaders have lower power (that arises through dependence), and thus non-leaders exhibit greater language accommodation than student leaders.

2. The language accommodation can propagate through the network, and specifically student leaders may have a greater reach of their influence by coordinating other people indirectly.

In this report, we present our preliminary evidence that supports Hypothesis 1. We have not yet implemented the graph structure to perform experiments that support Hypothesis 2.

We have implemented the coordination measure $C^m(b, a)$ proposed in [5], which is defined as follows:

$$C^m(b, a) = P(E^m_{u_b\to u_a}|E^m_{u_a}) - P(E^m_{u_b\to u_a}) \tag{1}$$

where b is the speaker that coordinates towards the target a, $E^m_{u_b\to u_a}$ is the event that the utterance of b exhibits a linguistic marker m in its reply to the utterance of a, and $E^m_{u_a}$ is the event that the utterance of a exhibits a marker a.

In a thread-based discussion forum like our MOOC dataset, it is hard to detect who is talking with whom, as discussed in our Assignment 2. Therefore, in this report we define the conversational exchange between b and a if b's post appears after a's post in the same thread. This assumption is limiting, but simple to implement.

Note that [5] employs the LIWC-derived categories for the lingustic marker m, where the categories include articles, auxiliary verbs, conjunctions, high-frequency, etc. (total 451 lexemes). We do not have access to these 451 lexemes yet, and thus in this report we only study the language coordination on the use of English articles ('a', 'an', 'the').

Figure 1 shows the language accommodation for two different groups (leaders and non-leaders). We have hand-annotated 8 student leaders and 8 non-leaders from the dataset (with a very high inter-annotator agreement rate). We then average their language accommodation towards/from other 3600 authors. We assume that the majority of the rest of the 3600 authors are non-leaders.

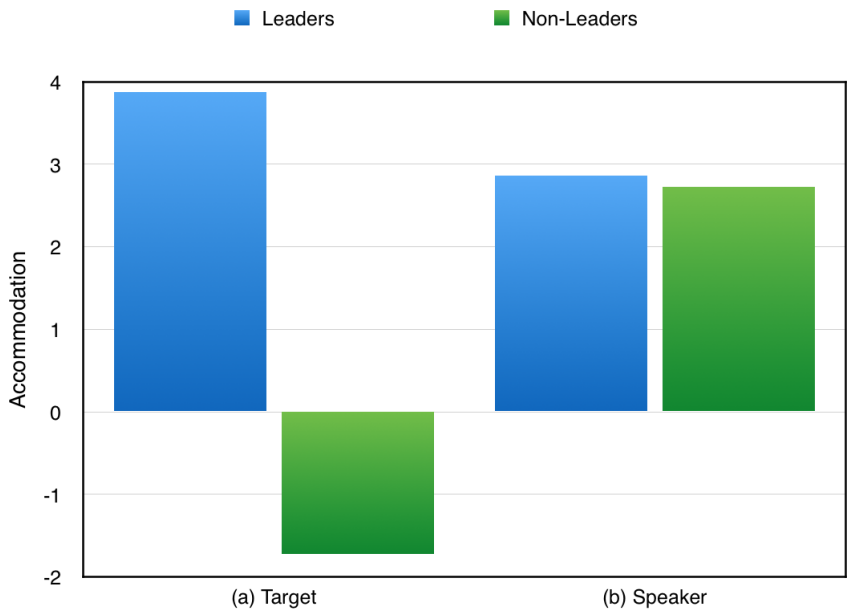

Figure 1: Language accommodation on the stylistic marker '**article**.' (a) Users coordinate more towards student leaders than towards non-leaders (Targets), supporting our Hypothesis 1. (b) On the other hand, we were not able to find a significant difference between leaders and non-leaders (Speakers) in language accommodation towards other people. The y-axis values are reported as percentages (i.e., multiplied by 100) for clarity.

The result above does support our Hypothesis 1, although the implications are rather limited. We acknowledge that the result we present in this section is not a rigorous claim for the following reasons:

- Only one linguistic marker (article) was considered due to time constraints. We need to perform the same analysis on other categories of linguistic style markers as well.
- We may need annotations for more leaders and non-leaders to get statistically significant results.
- Our definition of a conversational exchange might not always be valid.
- The corpus has not been totally cleaned (e.g. removal of the code snippets, etc.)

2.1.4 Future Work

We plan on addressing the issues we identified above to better support our Hypothesis 1. We will then implement the graph framework we proposed in Section 2.1.2, and test our Hypothesis 2. We can test Hypothesis 2 on the prediction task of student leaders, and use the plain coordination features as a baseline.

In other words, we expect that incorporating the information about language influence propagation in the features will improve the performance on the prediction task compared to the system learned only on the coordination features.

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:

$$\hat{M} = \underset{M \in \{M_I \cup M_S\}}{\operatorname{argmax}} \sum_{u \in U} \mathcal{L}(u|M) \tag{2}$$

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.

2.2.1 Preliminary Analysis

1. *Data Analysis and Conclusion*
   We observed that the number of posts by a student alone is very few to get good results by running topic modelling on it. The topic distribution generated doesn't give us a good insight into the topics being discussed by the student. Thus we also concluded that our approach of time-varying topic models would not work, as slicing up the data further will not improve results. We discarded the large number of posts that were "Anonymous", where we lost a large portion of the valuable data. This couldn't be avoided as there are a number of people who would comment as anonymous, and we can't identify the posts by them separately.

2. *Details of the Approach*

   We created topic models for "ETIENNE PAPEGNIES", one of the most active instructor/TA on the forum. He had 1024 posts. Then based on that we tried to predict the probability of each of the student's language having the same distribution as him.

   We studied most of the students who had 101 posts and above in the corpus. This is because we need large number of posts to perform topic modelling effectively.

   The details of the tools and parameters used for conducting the experiments:

   (a) We used MALLET natural language processing toolkit for performing topic modelling using an LDA (Latent Dirichlet Allocation) technique on the data.

Table 1: The number of users vs. number of posts

3311 users	1 to 10 posts
189 users	11 t0 20 posts
60 users	21 to 30 posts
23 users	31 to 40 posts
11 users	41 to 50 posts
10 users	51 to 60 posts
7 users	61 to 70 posts
8 users	71 to 80 posts
2 users	81 to 90 posts
2 users	91 to 100 posts
19 users	101 posts and above
Anonymous	4253 posts

(b) We used 5 topics as the data we have is very small. Also, since it is a very specific forum that we are working with, the lower number of topics gave us better results. A large number of topics affected the results adversely.

(c) We used both unigrams and bigrams as input to the LDA.

(d) The number of iterations were set to 50. Lower number of iterations gave poor results.

(e) The alpha was set to 0.01 and the beta was set to 0.01.

3. *Results of the Approach*
   The results we got were promising. We studied most of the students who had 101 posts and above in the corpus. The students who got a high score were almost always identified correctly as leaders. This is displayed in the excel sheet we have attached where we hand-picked the comments they have that are similar to instructors. There are varied examples of how the student leaders don't only solve problems but also tell other students what they found interesting and provide support and guidance by telling them if they are on the right track or not and also tell them what they can do to improve their learning. We also observed that they were slightly rude at times. This is in conjunction with what we had read during the course as to how people with power and knowledge tend to be less polite than others.

4. *Further Ideas for Improvement*
   There were also some wrong labels that were assigned to the students. This was because the student talked about the same topics as the instructors, but was mainly asking questions related to those topics. We can overcome this mislabelling by marking the posts as questions or answers base on simple techniques before passing them to the LDA. This way we can differentiate between people asking questions and people who are answering them. We also plan on running the experiments for other values of number of topics, number of iterations, alpha and beta to see if they improve the results.

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.

2.3.1 Preliminary Analysis

1. *Details of the Approach*
   Features were extracted using the LIWC corpus. We used all possible categories in the Internal Pennebaker LIWC2007 Dictionary, which included the following: word count, words per sentence, words greater than 6 letters, dictionary words, function words, pronouns (and its subcategories), articles, verbs (and its categories), adverbs, prepositions, conjunctions, negations, quantifiers, numbers, swear words, social processes, labels for family members, friends, and humans, affective processes, positive emotions, negative emotions, anxiety, anger, sadness, cognitive processes, insight, causation, discrepancy, tentativeness, certainty, inhibition, inclusiveness, exclusiveness, perceptual processes, seeing, hearing, feeling, biological processes, words to describe the body, health, sexuality, ingestion, relativity, motion, space, time, work, achievement, leisure, home, money, religion, death, assent, disfluencies, fillers, and various types of punctuation. The features were fed into various types of unsupervised clustering algorithms from Weka to find any patterns in the data. The results from these clusters were then compared to the role of the author of the post (instructor, student, or student leader).

2. *Results and Conclusion of the Approach*
   The following table lists the percentage of incorrectly clustered instances, compared to hand-labeled data. The 3-way comparison column is between three, equally-sized categories (student leaders, instructors, and regular students). The next column shows the accuracies of comparing instructors/leaders vs regular students. The instructors/leaders group here, which is half comprised of instructors and half student leaders, is the same size as the regular students group. The last column is the same two labels, but all of the data from the 3-way comparison is used such that the instructors/leaders category is twice as big as the regular students category.

Model	*3-way Comparison*	*2-way Comparison Even*	*2-way Comparison Uneven*
EM	58.33%	44.64%	50%
Filtered Clusterer	60.71%	35.71%	36.90%
Density Based Clusterer	59.52%	33.93%	35.71%
Simple K-Means	60.71%	35.71%	36.90%

   When the instructors are grouped together with the student leaders, the accuracy consistently improves across all models, regardless of the number of instances of posts from authors in each role. Therefore, future experiments look promising, and perhaps student leaders do tend to talk like instructors.

3. *Future Plans for Experiments*
   One of the next obvious steps would be to stem the words to see if that would make any difference in

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

the clustering and classification. We also have access to the General Inquirer dictionary, which we can use in addition to LIWC. It might also be beneficial to use more labeled data (target more student leaders in the corpus), so that we can create more detailed models. Now that we know that the student leaders could be similar to instructors in the way that they talk, perhaps we can create a model to classify the role of the writer and see if we can predict whether they are a student leader or not, in addition to figuring out which features are more indicative of the differences between the roles.

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.

[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.