Discovering Leadership Roles in Email Workgroups

Vitor R. Carvalho
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA
vitor@cs.cmu.edu

Wen Wu
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA
wenwu@cs.cmu.edu

William W. Cohen
Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA
wcohen@cs.cmu.edu

ABSTRACT

Email is a key communication tool for collaborative workgroups. In this paper, we investigate how team leadership roles can be inferred from a collection of email messages exchanged among team members. This task can be useful to monitor group leader's performance, as well as to study other aspects of work group dynamics. Using a large email collection with several workgroups whose leaders were previously defined, we demonstrate that leadership positions can be predicted by a combination of traffic-based and text-based email patterns. Traffic-based patterns consist of information patterns that can be extracted from the message headers, such as frequency counts, message thread position and whether the message was broadcast to the entire workgroup or not. Textual patterns are represented by the message's "*email speech acts*",i.e., semantic information with the sender's intent that can be automatically inferred by language usage. Using off-the-shelf learning algorithms, we obtained 96% accuracy and 88.2% in F-measure in predicting the leadership roles on 34 email-centered work groups.

1. INTRODUCTION

Email has become one of the most important means of communication in collaborative workgroups. In this paper, we investigate how email exchange patterns in work environments may reveal the underlying hierarchical structure of an organization. More specifically, we attempt to predict leadership roles in small work groups using information derived from all email exchanged within the group. For this purpose, we carefully analyzed a corpus of thousands of emails exchanged in a teamwork context[3]. In this corpus, 34 teams of 5 or 6 members worked together for several months on the same project, with leadership roles clearly assigned in the beginning.

Predicting the leadership role can be useful to keep track of the leader's performance, which in turn might affect several other aspects of the group's behavior. For instance, in groups where leadership is weak or not obviously assigned, some team members tend to naturally converge to a leadership role, and such converged leaders could be automatically detected. Many application scenarios can be imagined for a leadership predictor, particularly in evaluation of performance of teams (and team members) on collaborative projects or even in studying different styles of group leadership.

Our findings suggest that one can accurately predict group leading roles based on traffic-based patterns and on the textual contents of the messages. We were able to predict the correct leadership role in approximately 95% of the test cases. Additionally, using statistical tests we calculated what particular features are highly associated with leadership positions, revealing a surprising agreement with the common intuition on this problem.

2. DATASET: THE GSIA CORPUS

In our experiments, we used a large email collection called the GSIA corpus. The GSIA corpus[3] contains approximately 15,000 emails and was collected from a management course at Carnegie Mellon University. Only emails exchanged among course members were collected. In this course, 277 MBA students, organized in 50 teams of four to seven members, ran simulated companies in different market scenarios over a 14-week period. Because some teams had less than 20 emails logged, which we considered insufficient, only 34 teams, with an average of 6 members per team and a total of 12322 emails, were eventually considered in our study.

Every team had its president assigned in the beginning of the game, as well as other managerial roles, such as VP Marketing, VP Finance, VP Finance, etc. The president represented the highest position in the hierarchy of the team, but hierarchical relations among other roles were not clear. Therefore, we aimed to infer leadership only in terms of which member was the president in the team. In contrast with many related work, this fact allowed us to provide a very objective evaluation of leadership.

In general, this email dataset tends to be very task-oriented, with several instances of task delegation, negotiation, delivery of files and meeting arrangements. Since the teams were competing with one another, message exchange between two different teams is rare, whereas internal communication within members of the same team was very frequent. This corpus has interesting characteristics due to the fact that the teams worked mostly in isolation from all other teams. We observed that even though the majority of exchanged messages in all work teams were personal, or only delivered to part of the team members, there was a significant number of messages sent to all team members (a.k.a. broadcast messages).

3. TRAFFIC-BASED FEATURES

CEAS 2007 - Fourth Conference on Email and Anti-Spam, August 2-3, 2007, Mountain View, California USA

Motivated by the nature of the corpus, we divided emails into two types: *broadcast* and *non-broadcast* emails. Broadcast emails are messages sent by one team member and addressed to all other teammates. The non-broadcast emails are the ones not addressed to all team members. All features are extracted on a user basis, and normalized according to their class.

Table 1 (part a) lists the complete set of broadcast traffic-based features. For instance, *bcast_count* represents the number of broadcast emails sent by a user normalized by the total number of broadcast emails in her team. The feature *bcast_StartThread_count* expresses the normalized number of broadcast emails that start an email thread (message discussion). External addresses are email addresses from outside of the team. Thus, features such as *bcast_ExternalAddress_count* represent the normalized number of broadcast messages with at least one external recipient (i.e., it was sent to all group members in addition to least one outside email address). Similar features can be extracted for the non-broadcast messages; with the addition of two new types of features: the number of CC'ed messages and the number of received emails (see Table 1 (part b)).

In addition to the features listed in Table 1, we added the ranking of each feature as new features. The ranking of each feature indicates if the user is ranked first (_1), second (_2), last(_last) or one but last(_butlast) for the feature. For instance, *bcast_StartThread_count_1* indicates that the user send the most broadcast emails starting a new thread in the team; and *nonbcast_count_receive_last* reveals that this particular user received the smallest number of non-broadcast messages in the group.

(a) Broadcast Email Features
bcast_count
bcast_EndThread_count
bcast_StartThread_count
bcast_ExternalAddress_count
bcast_ExternalAddress_StartThread_count
bcast_NoExternalAddress_count
bcast_NoExternalAddress_StartThread_count
(b) Non-Broadcast Email Features
nonbcast_count_send
nonbcast_count_receive
nonbcast_count_cced
nonbcast_StartThread_count
nonbcast_EndThread_count
nonbcast_External_count
nonbcast_ExternalAddress_count
nonbcast_ExternalAddress_StartThread_count
nonbcast_NoExternalAddress_count
nonbcast_NoExternalAddress_StartThread_count
(c) Email Speech Act Features
[act]=request or commit or propose or meet or deliver
user_bcast_[act]_count
user_bcast_[act]_startingThread_count
user_bcast_[act]_endingThread_count
user_nonbcast_[act]_rcvd_count
user_nonbcast_[act]_sent_count
user_nonbcast_[act]_sent_startingThread_count
user_nonbcast_[act]_sent_endingThread_count

Table 1: Sets of Features

4. CONTENT-BASED FEATURES

In a previous work, Cohen et al.[2] used text classification methods to predict "email speech acts". Based on the ideas from Speech Act Theory [5] and guided by analysis of several email corpora, they defined a set of "email acts" (e.g., *Request*, *Deliver*, *Propose*, *Commit*) and then classified emails as containing or not a specific act. Cohen et al. [2] showed that machine learning algorithms can learn the proposed email-act categories reasonably well. It was also shown that there is an acceptable level of human agreement over the categories.

More recently, Carvalho and Cohen [1] released an open source package for *Email Speech Act* classification. We decided to use its predictions on five different acts (*Commit, Request, Deliver, Propose and Meeting*[1]) as textual features on the task of leadership prediction. The features are specified in Table 1(part c). For instance, the feature *user_bcast_[act]_count* expresses the number of broadcast messages of a particular user that contained a "Request Act" in its textual contents. The "email speech act" features were normalized in the same way of the broadcast and non-broadcast features. Similarly, we also used the associated ranking features, i.e., first (_1), second (_2), last (_last) and one but last (_butlast) on the "email speech act" feature set.

5. LEADERSHIP ROLE PREDICTION RESULTS

Having extracted all features, the task was then formulated as a binary classification problem. We used an off-the-shelf linear SVM[2] classifier to predict the leadership roles. Experiments were performed in 10-fold cross-validation setting, where all members of the same team were kept in the same validation set (either train or test).

Using the entire feature collection (*All Features = Bcast + NonBcast + SpeechAct*) from Table 1, the classifier reached approximately 94.5% accuracy and about 83.5% of F1-measure[3]. We then investigated how different types of feature contribute to this task. Figure 1 illustrates accuracy and F1-measure values for different feature sets.

The last column (*All Features*) of Figure 1 refers to the situation where all textual and traffic features are used, i.e., *broadcast* and *non-broadcast* and *speech act* features. The first column (*baseline*) uses two features only, *bcast_count_1* and *nonbcast_sent_count_1*. This baseline simulates the criteria "who sent the largest number of (broadcast and non-broadcast) messages". It is interesting to notice that, just by using the baseline, we can make the correct prediction in 90% of the cases, indicating that the presidents do tend to send more messages than the other team members.

The second column illustrates the performance when only non-broadcast features are used. Similarly, the third column shows the same number when only the broadcast features are used. In the fourth column, broadcast and non-broadcast features were used. Figure 1 clearly indicates that broadcast features are more informative than non-broadcast features

[1]Detailed descriptions of these "acts" can be found in [1].

[2]We used the LIBSVM library with default parameters.

[3]The harmonic mean of precision and recall

for this task. This is somewhat expected since group leaders are more expected to motivate, communicate good/bad news, deliver performance updates, etc. to the entire group. When broadcast and non-broadcast features are combined, the performance is considerably better than the baseline. It is also interesting that adding the textual information (speech act features) does improve the prediction results of the system. In the next section, we provide a more detailed analysis of which features are the most associated with the leadership concept in this dataset.

In this binary classification task, there were no guarantees that all workgroups would have exactly one team member as leader. We then added a post-processing step to make sure that all teams would have exactly one leader — the member with highest confidence from the classifier. With this change, testing with the *All Features* setting achieved 96% of accuracy and an F1-measure of 0.882. Table 2 shows the confusion matrix on this prediction: it correctly predicts the president in 30 out of 34 groups.

	Predicted	
	Leader	Not Leader
True Leader	**30**	4
True Not-Leader	4	**160**

Table 2: Final Confusion matrix - *All Features*.

6. DISCUSSION

In order to quantify the importance of the different features, we applied the Chi-Square independence test. The idea is measuring how independent the feature presence and the label(president or not) are, when considered as random variables. The higher the score, the less likely the feature is to be independent from the label. Table 3 shows the top 10 ranked features with the corresponding Chi-Square scores (χ^2). As expected, Table 3 reveals that leadership is closely associated to the largest number of messages sent, broadcast and nonbroadcast, including or not external addresses. More specifically, the top two features (broadcast messages starting or ending a thread of conversation) are very strong leadership features; which agrees with the intuition that leaders do tend to start (and end) new discussions, proposals, update messages, requests, etc. in the group.

The 6th most meaningful traffic feature (nonbcast_count_receive_1) shows that the president tends to be privately addressed in non-broadcast emails. Also, the presence of the nonbcast_count_cced_1 feature in the list reveals that team members tend to add the president to the recipient list when dealing with other teammates[4].

The speech act feature list in Table 3 reveals that leaders are frequently associated with meeting related messages; particularly broadcast ones starting new threads of conversation (possibly proposing a new group meeting). Another interesting feature in the list is user_nonbcast_deliver_rcvd_count_1, indicating that the team leader tends to receive a lot of nonbroadcast messages delivering some information. This is the only receiving feature in the list. Even though the speech acts features improved the overall results for the task, the way they were formulated seems to be largely overlapping with the traffic features - this would possibly explain the small gains in performance in Figure 1.

[4]This agrees with the intuition that making the boss aware of a request often increases the chances of having it completed quickly.

(a) Top Bcast+NonBcast Features	χ^2
bcast_endThread_count_1	2.219
bcast_StartThread_count_1	2.209
nonbcast_count_1	2.209
nonbcast_NoExternalAddress_count_1	2.209
bcast_count_1	2.113
nonbcast_count_receive_1	2.098
bcast_NoExternalAddress_StartThread_count_1	1.913
nonbcast_endThread_count_1	1.913
bcast_NoExternalAddress_count_1	1.804
bcast_ExternalAddress_StartThread_count_1	1.776
nonbcast_NoExternalAddress_StartThread_count_1	1.776
bcast_ExternalAddress_count_1	1.702
nonbcast_count_send_1	1.624
nonbcast_NoExternalAddress_count_1	1.624
nonbcast_ExternalAddress_StartThread_count_1	1.541
nonbcast_StartThread_count_1	1.541
bcast_ExternalAddress_StartThread_count	1.330
nonbcast_external_count_1	1.319
nonbcast_count_cced_1	0.872
nonbcast_count_send_2	0.654
(b) Top Textual Features	χ^2
user_bcast_meet_count_1	2.261
user_bcast_meet_startingThread_count_1	2.154
user_bcast_deliver_startingThread_count_1	2.098
user_nonbcast_deliver_rcvd_count_1	2.098
user_bcast_deliver_count_1	1.997
user_bcast_request_count_1	1.846
user_bcast_request_startingThread_count_1	1.846
user_nonbcast_meet_sent_count_1	1.846
user_nonbcast_request_sent_count_1	1.846
user_nonbcast_deliver_sent_count_1	1.776

Table 3: Top ranked features – according to Chi-Squared (χ^2) scores

Another weakness in the current approach is the fact that this particular dataset can be considered "too clean" if compared to real email corpora. In fact, applying the same techniques to other email collections require previous knowledge of team structure, which is not always available and/or well defined. However, the main motivation for using this dataset was the fact that it had team members and leadership positions clearly assigned in the beginning of the data collection, which allowed unambiguous evaluation of the leadership prediction task.

7. RELATED WORK AND CONCLUSIONS

In [4], Leuski focused on one particular aspect of the email problem: detecting people's roles. The author claims that a single person can "play" different "personas" and these roles are reflected in the content of all her communications with the outside world and thus in her email messages as well. An SVM classifier is trained for every speech act on a 500 message dataset. There are five people within the group who are "professor, head of the research group", "graduate stu-

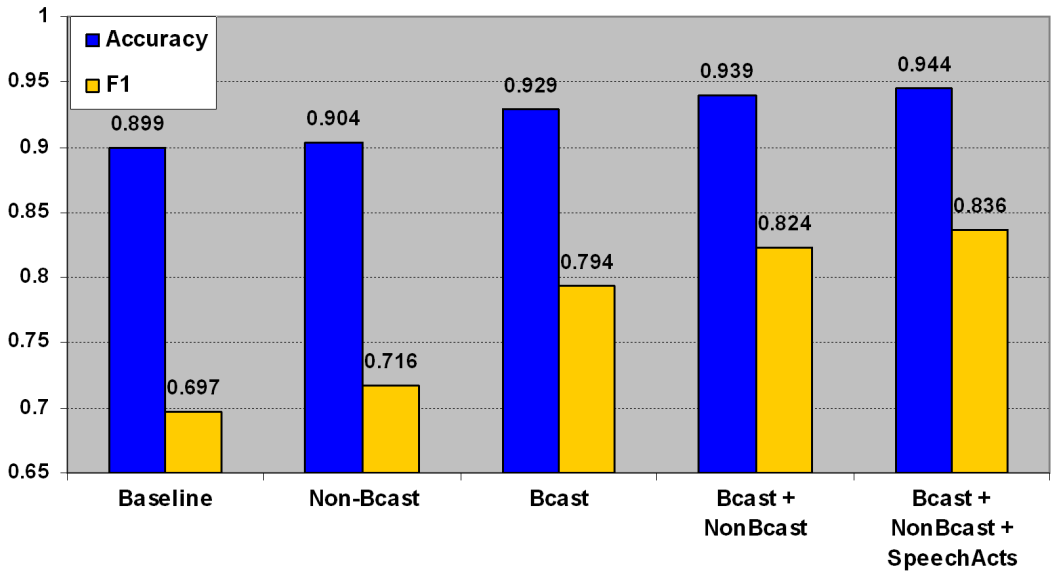

Figure 1: Comparison of Accuracy and F1-Measure with different feature sets.

dent", "secretary", "researcher", and "programmer". The experiments achieve seem to achieve good accuracy levels, however the results presented in [4] were derived from a very limited amount of data and it is hard to know if these results would scale well to larger datasets.

Tyler et al.[6] proposed a *betweenness centrality algorithm* for the automatic identification of communities of practice from email logs of a large organization. They also investigated how to identify leadership roles in their corpus, but the evaluation was more subjective than the one performed here.

In this paper we presented a new method for predicting leadership roles from email collections by using textual ("email speech acts") and traffic-related (broadcast and non-broadcast messages) features extracted from email messages. Results indicate that these features are very good predictors of leadership positions in email-centered workgroups. A detailed analysis revealed that broadcast messages are better leadership indicators than the non-broadcast ones; and that textual features can help in predicting the leading positions in a group. We reported accuracies of approximately 96% for the leadership prediction task in a large email collection with 34 different workgroups.

8. REFERENCES

[1] V. R. Carvalho and W. W. Cohen. Improving email speech act analysis via n-gram selection. In *Proceedings of the HLT/NAACL 2006 – ACTS Workshop*, New York City, NY, 2006.

[2] W. W. Cohen, V. R. Carvalho, and T. M. Mitchell. Learning to classify email into "speech acts". In *Proceedings of EMNLP 2004*, pages 309–316, Barcelona, Spain, July 2004.

[3] R. Kraut, S. Fussell, F. Lerch, and A. Espinosa. Coordination in teams: Evidence from a simulated management game. To appear in the Journal of Organizational Behavior, In submission.

[4] A. Leusky. Email is a stage: Discovering people roles from email archives. In *ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 2004.

[5] J. R. Searle. *A taxonomy of illocutionary acts*. In K. Gunderson (Ed.), Language, Mind and Knowledge. Minneapolis: University sof Minnesota Press, 1975.

[6] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman. Email as spectroscopy: Automated discovery of community structure within organizations. In *Communities and Technologies*, 2003.