Inferring Organizational Titles in Online Communication

Galileo Mark S. Namata Jr.[1], Lise Getoor[1], and Christopher P. Diehl[2]

[1] Dept. of Computer Science/UMIACS, Univ. of Maryland, College Park, MD 20742
[2] Johns Hopkins Applied Physics Lab., 11100 Johns Hopkins Rd., Laurel, MD 20723

1 Introduction

There is increasing interest in the storage, retrieval, and analysis of email communications. One active area of research focuses on the inference of properties of the underlying social network giving rise to the email communications[1, 2]. Email communication between individuals implies some type of relationship, whether it is formal, such as a manager-employee relationship, or informal, such as friendship relationships. Understanding the nature of these observed relationships can be problematic given there is a shared context among the individuals that isn't necessarily communicated. This provides a challenge for analysts that wish to explore and understand email archives for legal or historical research

In this abstract, we focus on a specific subproblem of identifying the hierarchy of a social network in an email archive. In particular, we focus on and define the problem of inferring a formal reflection of an organizational hierarchy, the formal title of an individual, within the underlying social network. We present a new dataset, to use in conjunction with the original Enron dataset, for studying the formal organizational structure underlying an email archive. We also provide preliminary results from the classification of individuals to broad titles in the organization, relying only on simple traffic statistics.

2 Enron Hierarchy Dataset

One major impediment to research in hierarchy inference from email archives is the lack of a publicly available dataset providing email traffic from a structured organization along with documentation of that structure. Therefore, we present a dataset[3] for use with the Enron email dataset[3], with a particular focus on identifying the identities and titles of the individuals from whose accounts the Enron email dataset was generated. Using various forms of the dataset[4, 5, 6], along with documents related to the Enron trial, we identified the individuals whose accounts compose the Enron dataset, as well as the email addresses, titles and company groups for 124 of them. We also provide a mapping to six broad titles from similar titles i.e.: VP (Vice President) for VP of Finance.

[3] Available from http://www.cs.umd.edu/projects/linqs.

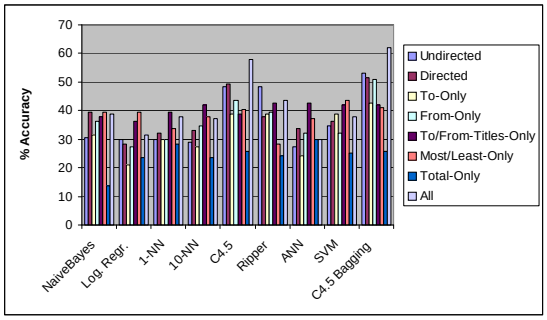

Fig. 1. Summary of Broad Title Classification Accuracy

3 Problem

In order to identify the underlying social network hierarchy, we focus on a formal representation of this hierarchy, the formal titles of individuals. Specifically, we focus on the problem of mapping the set of actors (people who send/receive emails in an email collection) to a set of formal titles within the organization.

In our experiments, we classify the 124 individuals in our extended dataset to six broad titles using various classifiers[7] processing simple statistics derived from the relevant traffic. We use three types of traffic statistics: undirected (# of emails sent and received), directed(# of emails sent, # of emails received) and aggregate (# of emails sent/received to labeled individuals of a given broad title). The goal is to identify which set of statistics yields the best performance.

Fig. 1 shows a summary of the results from different classifiers using our various traffic models. Given the class distribution of the six target broad titles has an average random classifier performance of 20% accuracy, the results are promising. In general, our classifiers outperformed the random baseline by a statistically significant margin. Using our undirected model, we received an accuracy of 53.2%, over twice as well as random. We note that although the undirected model lacks traffic direction information, it performed comparably with the directed model. The same is true in variations where we only used one direction of the traffic. This implies that we may be able to classify individuals without having all their email communications. Finally, we note that when we use our aggregate model with the other two models, we are able to reach an accuracy of over 62%.

It is also interesting to examine the confusion matrix, in Fig. 2, to see where the misclassifications are occurring. Of note is the correlation between the misclassifications of individuals to titles. The misclassification of a title seems to occur mainly with titles close to the correct title in the hierarchy. For example, *DIR* is mainly misclassified with its immediate superior, *VP*, and subordinate, *MGR*. Similarly, *ASSOC* is most misclassified as *MGR*. This trend is consistent

	CO	VP	SPEC	DIR	MGR	ASSOC
CO	8	3	0	0	0	0
VP	0	21	0	1	1	1
SPEC	0	0	11	4	1	1
DIR	1	4	2	20	5	0
MGR	2	3	1	12	8	2
ASSOC	0	0	0	0	3	9

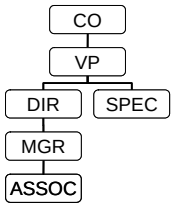

Fig. 2. Confusion matrix for broad titles correspond to title hierarchy.

among all the titles. This implies that our approach using traffic statistics might be able to reconstruct levels in the overall hierarchy.

4 Conclusion

Identifying the hierarchy of the underlying social network is an important problem which aids in the exploration and understanding of organizational email archives. We investigate a component of this larger problem, mapping individuals to their formal titles in the organization, using only simple traffic statistics and present preliminary experimental results. In future work, we would like to make greater use of commonly used social network analysis measures such as centrality and equivalence in our classifiers, as well as use the content of the email messages. We are also interested in trying to classify other relationships, such as friendships or direct report relationships. Finally, we are interested in the temporal aspects of the hierarchy, specifically being able to detect how and when the organizational structure changes in an archive.

References

[1] Diehl, C., Getoor, L., Namata, G.: Name reference resolution in organizational email archives. In: 6th SIAM Conference on Data Mining, (Bethesda, Maryland)
[2] Corrada-Emmanuel, A., McCallum, A., Wang, X.: Topic and role discovery in social networks. In: IJCAI. (2005)
[3] Klimt, B., Yang, Y.: Introducing the Enron corpus. In: Conference on Email and Anti-Spam. (2004)
[4] Adibi, J.: Enron dataset (2005) http://www.isi.edu/ adibi/Enron/Enron.htm.
[5] Fiore, A.: UC Berkeley Enron email analysis (2005) http://bailando.sims.berkeley.edu/enron_email.html.
[6] Corrada-Emmanuel, A.: Enron email dataset research (2004) http://ciir.cs.umass.edu/~corrada/enron.
[7] Witten, I., Frank, E.: Data Mining: Practical machine learning tools and techniques. Volume 2nd Edition. Morgan Kaufmann (2005)