Inferring Formal Titles in Organizational Email Archives

Galileo Mark S. Namata Jr. NAMATAG@CS.UMD.EDU
Department of Computer Science/UMIACS, University of Maryland, College Park, MD 20742, USA

Lise Getoor GETOOR@CS.UMD.EDU
Department of Computer Science/UMIACS, University of Maryland, College Park, MD 20742, USA

Christopher P. Diehl CHRIS.DIEHL@JHUAPL.EDU
Johns Hopkins Applied Physics Laboratory, 11100 Johns Hopkins Road, Laurel, MD 20723, USA

Abstract

In the social networks of large groups of people, such as companies and organizations, formal hierarchies with titles and line of authority are established to define the responsibilities and order of power within that group. Although this information may be readily available for individuals within the group, the context this hierarchy provides in communications is often not available to those outside the organization. In this paper, we define the problem of inferring this formal hierarchy in the context of organizational email archives. We describe a new dataset which extends the widely used Enron email dataset with information about the formal organization hierarchy. We then provide some preliminary results on the problem of classifying individuals to their formal titles using standard classification algorithms.

1. Introduction

Email communications generate an estimated annual 400,000 terabytes of information each year, second only to the telephone as the main form of electronic communication (Lyman & Varian, 2003). As such a major form of communication, there is increasing interest in the storage, retrieval, and analysis of email traffic.

One current area of research is the inference of properties of the underlying social network in email communications. Email communications between individuals imply a relationship between those individuals, whether it is formal, such as a manager-employee relationship, or informal, such as friendship relationships. Although these relationships are known within the social network, they are not readily available outside the group. This provides a challenge for analysts to explore email archives for legal or historical research.

An example of this challenge is in the field of reference resolution. A common example of this is when two individuals refer to a third individual by a first name or nickname in a communication. Consider the example:

"Taffy, Is Mark available to meet tomorrow? Susan"

From the flow of the conversation, it is clear that both individuals know who Mark is. However, for someone outside the group, who does not know that Taffy is the secretary of a Mark Taylor, the reference is meaningless.

A similar problem can be seen in the identification of key individuals in a social network. If we are interested in identifying who the manager of a particular group is and we do not have explicit documentation of this information, it may take analysis of thousands of emails to infer the information. On the other hand, someone within that group would be able to easily identify that person with the shared knowledge of that organization.

In this paper, we focus on a specific sub problem of identifying the hierarchy of a social network in an email archive. In particular, we focus on and define the problem of inferring the formal title of an individual within the underlying social network of an organizational email archive. We also present a new dataset, to use in conjunction with the original Enron dataset, for use in studying the formal hierarchies underlying

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.

an email archive. We provide preliminary results from one approach to the classification of individuals to a broad title in the organization, relying only on simple traffic statistics.

2. Problem Description

Let $\mathcal{A} = \{a_j\}$ be the set of actors (people who send and receive email, in an email collection and let $\mathcal{T} = \{t_j\}$ be the set of formal titles used within the underlying social network of an organizational email archive. The set $\mathcal{T}$ can either be titles specific to a single individual within the social network, or can be generalized to a broad title of a group. For example, Vice President of Finance and Vice President of Operations can be two separate titles or can be generalized to Vice President. The objective of formal title inference is to construct a mapping from the set of known actors $\{a_j\}$ in A to a single title $\{t_j\}$ in $\mathcal{T}$.

Note that at this point, we only define the problem of mapping an individual to a single title. Although mapping to a single title will be the focus of our paper, the problem can be generalized to mapping a name to multiple titles. Such mappings may be necessary in cases where a single individual might have multiple formal titles in a company. Moreover, this addresses the temporal aspect of an organization's social networks, where titles of individuals may change over time.

Email archives can be classified as a collection of unstructured and structured data. The unstructured data, or content, of the email communications include the body and attachments of the emails. Although we are able to discover formal titles for individuals using this method, this approach often does not provide title information for all individuals. Although in some cases title information is included in email communications, such as in email signatures, in general, this information is already known within the group and is unnecessary to repeat. The structured part, or metadata, of emails, consists of the sender, recipients and date of an email message. It has been shown that structured data shows communication patterns that may be exploited to identify relationships between individuals, without being dependent on the content of emails (Diehl et al., 2006; Tyler et al., 2005). The structured data of email archives will be the focus of our paper.

3. Enron Hierarchy Dataset

One major difficulty with analysis of hierarchies from email archives is the current lack of a publicly available dataset which contains a large amount of email traffic from a structured organization, as well as formal documentation on that structure. In the next section, we describe a dataset for use with the Enron email dataset, with a particular focus on identifying the identities and titles of the individuals from whose accounts the Enron email dataset was generated.

3.1. Enron Dataset

The release of the Enron email dataset in 2003, as a part of the Federal Energy Regulatory Commission (FERC) investigation into Enron's accounting, provided researchers a unique glimpse of email communications inside a major corporation. Consisting of 619,446 messages from the email accounts of over 150 individuals, the Enron dataset is now released under several forms (Klimt & Yang, 2004b). The original, raw format was first publicly released by CMU (Cohen, 2004). The data has been normalized and converted to a MySQL database by both USC and UC Berkeley (Adibi, 2005; Fiore, 2005). For our experiments, we mainly used the UC Berkeley version of the dataset since that version provides traffic counts between email addresses. We also used the raw format provided by CMU to identify email addresses discussed in the next section.

3.2. Enron Archive Base

Although there are thousands of email addresses in the Enron archive, belonging to an equally large number of individuals and groups within Enron, the communications for a majority of those are not fully observable. That is, the communication for the majority of email addresses only consist of the email messages sent to and from the roughly 151 individuals whose accounts were subpoenaed in the investigation. Thus the only fully observable email addresses, the addresses for which we can claim to hold a majority of to and from traffic, are those which belong to this subset of individuals. Since our research is focused on identification of titles using the structured archive data, we begin with the identification of this subgroup.

For brevity, in the rest of this paper, we shall refer to the group of individuals, whose email accounts were gathered for the Enron dataset, simply as the archive base.

3.3. Employee Information

Despite the uniqueness of the archive base, there are few resources on the identity of the individuals and what their positions were within Enron. The first, and most cited, resource is Corrada-Emmanuel's work mapping the 150 email account folders from the

raw dataset to a set of email addresses (Corrada-Emmanuel, 2004). Our analysis of the Enron email system, however, reveals that there are several email address formats present in the Enron dataset that Corrada-Emmanuel's mapping does not include. For example, for Phillip Allen and his folder, *allen-p*, we were able to find three more email addresses. One particular address, *pallen@enron.com*, contains 176 received email messages from 71 different email addresses, a significant omission for any analysis that relies on exchange counts between individuals. Another issue is that the email account directories do not always correspond to a single Enron employee. First, we found that at least three directories contain emails from multiple individuals. The directory, *mcconnell-m*, for example, contains emails for both Mark McConnell and Michael McConnell. Similarly, *fischer-m*, contains emails for Mark Fischer and Mary Fischer. We also have a very unusual case where not only do two individuals, Mark Dana Davis and Dana Davis, share the same email account folder, but they, for a time, both used the email address *dana.davis@enron.com*. Next, we note that multiple directories seem to belong to the same individual. For example, *whalley-g* and *whalley-l*, both seem to belong to Lawrence G. Whalley.

A second resource is the ISI Enron employee status data (Adibi, 2005). The provided spreadsheet maps names of 151 Enron employees to broad titles within Enron. Also, for a few individuals, there are additional comments regarding the group and specific titles of the individual. The list of employees in this dataset does not correspond directly to the list of employees whose accounts were used for the dataset, but do cover a majority of them. The set is largely incomplete, however, with 40 of the 151 individuals listed simply as *Employee* and 29 listed as *N/A*.

Since our experiments require a list of individuals to their title within the company, as well as a more accurate list of email addresses, we need to gather additional information about the individuals of the archive base.

3.4. Archive Base Identification

We begin with the assumption that each folder in the raw data set corresponds to a single individual. Though, as we mentioned above, this is not always the case, it is true for a majority of the account directories in the dataset. We then map the folders to names found in the emails of that folder, with the assumption that the name of folder is of the format <*lastname*>-<*firstinitial*>. Next, we modified the list for the special cases we discussed, where there is not a one to one mapping between individual and folder. The result is a list of the account folder, first name, last name and middle initial, where applicable, for 151 individuals. We also provide a common name for cases where an individual does not normally use their proper name in the collection, i.e.: Harry Arora for Harpreet S. Arora.

Abbreviation	*Description*
CO	Corporate Officer i.e.: CEO, CFO, CRO
VP	Vice President
DIR	Director
SPEC	Specialist
MGR	Manager
ASSOC	Associate

Table 1. Broad titles in the archive base.

3.5. Employee Titles

In order to get the title information for the archive base, we manually examined resources from the Enron legal proceedings and used references for these individuals found in Google. Moreover, we went through email messages for these individuals to find any reference to their titles including, but not limited to, title information listed in their signature, references to promotions, formal email introductions, and correspondences with the human resources department. We were also able to acquire and use an internal Enron phone list which lists the names, email addresses, titles and internal phone numbers for the Enron Wholesale Services (EWS) group during the archive's time period. In all, we were able to assign a broad title for 124 of the archive base individuals.

3.6. Employee Emails

For the email addresses of the archive base, we focus only on the Enron domain email addresses of these individuals. To begin with, within Enron, we identified the major email address generation schemes used to assign email addresses to individuals. We find that most of the names use some combination of names and initials, with some schemes using the initials of a specific group within Enron as a subdomain (i.e.: <*address*>*@ees.enron.com* for employees of Enron Energy Services). Five of the major schemes are:

- <*firstname*>.<*lastname*>*@enron.com*
- <*lastname*>.<*firstname*>*@enron.com*
- <*middleinitial*>..<*lastname*>*@enron.com*
- <*firstinitial*><*lastname*>*@enron.com*

- *<firstinitial><lastname>@<enrongroup>.enron.com*

With these patterns, we were able to generate email addresses for the archive base using the names of the archive base individuals. We then compared the generated email addresses to the list of all known email addresses in the email archive using the Berkeley database version of the Enron dataset. Finally, we filtered the email addresses by counting the number of times those email addresses appeared in the email account folder for the specific individual, as well as the similarity of the name part of the email address to the individual's name. Email addresses that appeared frequently as the recipient of an email in the inbox were accepted. Also, email addresses that appeared as the sender in any of the sent folders were accepted. Note that since this method of filtering is highly dependent on the assumption that each folder belongs to only one individual, it will not work for the special cases discussed before. Thus, we do not perform the filtering step on those instances but instead manually verify their email addresses.

4. Traffic Models

Given the Enron dataset and the title information for the Enron archive base, we now present an initial attempt to classify actors to their broad titles using the structured part of their communication.

4.1. Classifiers

As a first attempt for this classification, we tried different categories of classifiers to identify which approach type performs the best. For this, we used the implementations of nine classifiers in the Weka tool (Witten & Frank, 2005). Since the performance of these classifiers can be affected by the parameters used, we present the results for these algorithms using only the best performing parameters we found during the experiments. This does not guarantee that the classifiers were configured to perform optimally, but does provide a good estimate of their performance.

4.2. Traffic Statistics

The focus of our experiment will be classification of individuals to their formal titles using the structured data of email messages. For our experiments, the structured data we will concentrate on are the counts of the traffic to and from the archive base. We use three main variations of the count: undirected, directed and aggregate. The goal is to identify which set of statistics yield the best performance. We discuss these variations below and provide a chart for reference. Finally, we use the broad titles we created in the dataset as the target attribute of our classifiers. The distribution and description of these titles are shown in Figure 1.

Classifier	*Abbreviation*
Naive Bayes	NaiveBayes
Logistic Regression	Log. Regr.
1-Nearest Neighbor	1-NN
10-Nearest Neighbor	10-NN
C4.5 Decision Tree	C4.5
Ripper	Ripper
ANN	ANN
SVM	SVM
Bagging with C4.5	C4.5 Bagging

Table 2. Classifier Algorithms and the abbreviations used in our charts.

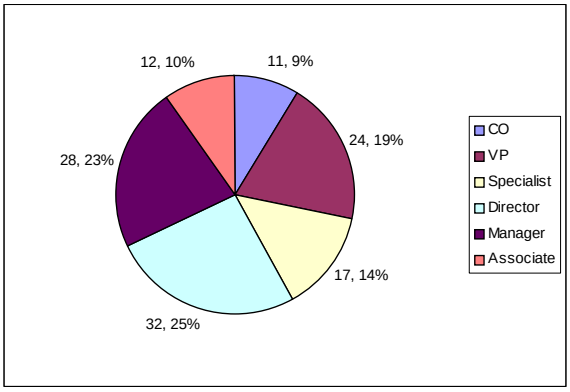

Figure 1. Class distribution of archive base broad titles.

4.2.1. Undirected Statistics

The undirected variation of the structured data is simply the overall count between two individuals. For example, the attribute *un_beck-s* for Paul Y'Barbo is number of email messages received from and sent to Sally Beck. Note that this means that *un_beck-s* for Paul Y'Barbo is the same value as *un_barbo-p* is for Sally Beck.

4.2.2. Directed Statistics

The directed version separates the traffic statistics between individuals based on direction. As an example, in the case mentioned above, *to_beck-s* refers to the number of emails Paul Y'Barbo sent to Sally Beck. Similarly, *from_beck-s* refers to the number of emails Paul Y'Barbo received from Sally Beck.

We are interested in how including directionality effects overall performance. In addition, because in some cases we may only have from or to traffic, we examine using each traffic direction statistics in isolation. Thus, aside from the variation which includes both the *to* and *from* values, we also create models base on the *to* counts and another with only the *from* counts.

4.2.3. Aggregate Statistics

As the name implies, for aggregate statistics we are interested in moving away from traffic counts of individuals and instead focusing on traffic between groups of people. The first, and simplest, are simple counts of the overall traffic that an individual has generated. This includes the total number of email sent, the number of emails received, the number of emails received as *CC*, and the number of emails received as *BCC*.

The second variation focuses on the traffic counts between groups of individuals based on the six broad titles we are using. We define the total number and percentage of an individual's email sent to and from a title. For example, *to_CO* and *from_CO* specifies the amount sent by individuals with the title of CO to and from an individual. We also have *to_CO%* and *from_CO%* as percentage of these values to the overall traffic of an individual.

The last variation we will use is the mapping of the counts to each title to four categorical attributes, *most-to*, *most-from*, *least-to*, and *least-from*. These attributes contain the title which received the most or least of a given traffic. For example, individuals who emailed *CO* more than any other title group have the *most-to* attribute set to *CO*.

5. Evaluation

We evaluate the performance of our classifiers and datasets in terms of their overall accuracy in classifying an individual to their broad title using leaveone-out cross validation. We compare this accuracy to the accuracy of a random classifier. We will also present the classbyclass true positive rate (TPR), false positive rate (FPR), precision (P), recall (R) and F-measure (F-1) values for the top performing algorithm, as well as discuss patterns in the misclassifications shown in the confusion matrix.

5.1. Classifier Performance

Figure 2 shows a summary of the results from different classifiers using the different categories of traffic statistics. The results from our evaluation of the classifiers, using the undirected version of the email metadata,

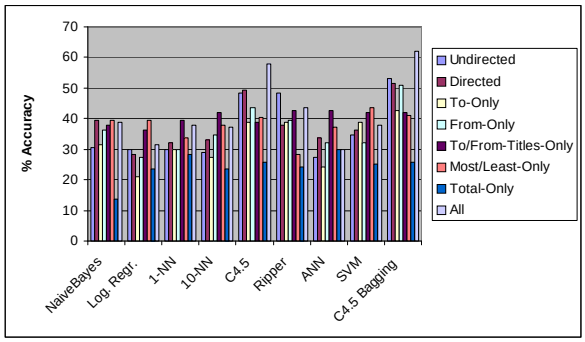

Figure 2. Broad title classification with aggregate traffic.

were very promising. Given the class distribution of the six target broad titles, the average performance of a random classifier is 20%. Our worst performing classifier, ANN, still performed statistically better than random at 27.5%. More notably, our best performing classifier, Bagging with the C4.5 Decision Trees, performed more than twice as well as random at 53.2% accuracy. Given this is only a baseline for these classifiers, with additional optimization possible in their parameters, these results are very encouraging.

5.2. Undirected vs. Directed

We expected directionality to improve the performance of our classifier, with the hypothesis that email communications to certain high ranking individuals were more significant than receiving emails. We found, however, that in most cases, classifiers using only undirected traffic performed just as well as directed traffic. Moreover, our best performing classifier actually performed better with undirected traffic statistics than with the directed traffic statistics.

Next we looked at using only a single direction of traffic, either to or from. In general, the performance of classifiers having only one direction of the traffic still resulted in comparable performance to classifiers using both directions. In four of our classifiers, the performance of having only one direction was within 1% of the performance with both directions. Also, looking again at our best performing algorithm, we see that using only the *From* part of our traffic, we achieve around 51% accuracy.

5.3. Aggregate

For readability, given the number of classifiers and variation we are using, as well as the fact that the classifiers performed similarly between the different versions of traffic, we focus our discussion on this section

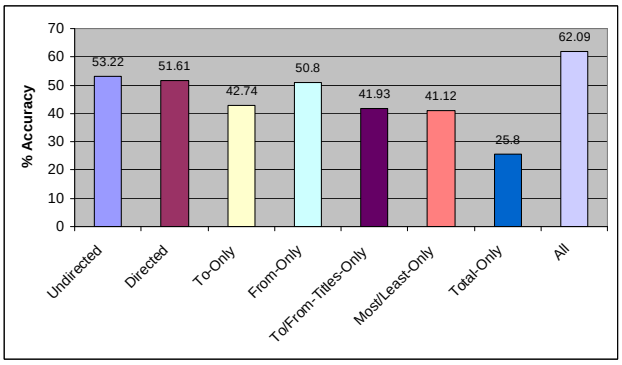

Figure 3. Broad title classification using Bagging with C4.5.

Title	*TPR*	*FPR*	*P*	*R*	*F-1*
MGR	0.321	0.083	0.529	0.321	0.4
VP	0.792	0.11	0.633	0.792	0.704
DIR	0.656	0.174	0.568	0.656	0.609
SPEC	0.647	0.037	0.733	0.647	0.688
ASSOC	0.75	0.036	0.692	0.75	0.72
CO	0.727	0.035	0.667	0.727	0.696

Table 3. Bagging with C4.5 performance by title

to the results to the best performing classifier, Bagging with C4.5 Decision Tree.

Figure 3 shows the results for this class using aggregate traffic statistics. First, we observe that with only the information of aggregate (the total, number of messages set, etc.), we can achieve over 40% accuracy, twice the performance of a random classifier. In the case of the *Most/LeastOnly* variation, this was achieved using only four categorical attributes, compared to the 124 continuous attributes used by *ToOnly*, which performed only slightly better. Finally, we observe that when we use these aggregate attributes with both the directed and undirected individual traffic, we were able to reach an accuracy of over 62%, roughly a 10% improvement over our best performance using only undirected.

5.4. Performance by Title

We now examine the results of the classifications, more closely focusing on where we are misclassifying. Below are the class by class values of true positive rate, false positive rates, precision, recall and f-measure values for Bagging with C4.5 Decision Tree shown in Table 3.

In general, the classifiers are performing the best on the highest and lowest ranked titles, namely *CO* and *VP*, on one end, and *ASSOC* on the other. Although this might be attributable to different email practices among the titles, we believe this may also be due to the fact that within Enron, these three titles were the least volatile groups. It has been documented that Enron had a 20% attrition rate due to promotions (J. Byrne, 2002). Since we are only classifying to one title over the whole time period, the misclassification of an individual may be due to the fact that that individual might have held the other position for a certain time period. In the case of Associates, *ASSOC*, the position is an entry level position and, as such, these individuals might not be misclassified simply since they have not had the opportunity to hold other positions. Similarly, *CO* and *VP* might be classified well since there is little promotion in these groups.

We also note the poor performance of our classifier on the position of *MGR*, Manager. Upon closer examination of the data, we found that there are certain individuals that were often misclassified. Investigation of this phenomenon reveals that this is especially true for two individuals, Mark Haedicke and Paul Allen. Although we listed these individuals as Managers, additional research revealed that the official title of these individuals is Managing Director. In the Enron hierarchy, the title of Managing Director is closer to that of Vice President than to a standard Manager.

We next look at the confusion matrix shown in Figure 4 of our classifiers to see where the misclassifications are occurring. These results are consistent with what we found when we looked at the classbyclass performance. Of note, however, is the correlation between the misclassifications of individuals to titles. The misclassification of a title seems to occur mainly with titles close to correct title in the hierarchy. In Figure 4, for example, *DIR* is mainly misclassified with its immediate subordinate, *VP*. Similarly, *ASSOC* is most misclassified as *MGR*. This trend is consistent among all the titles. This implies that our approach using structured data might be able to reconstruct levels in the overall hierarchy.

6. EWS Phone List Dataset

Given our performance in classifying individuals to their titles, despite missing certain directions of traffic, we wanted to evaluate the performance of our best performing classifier on a larger dataset. For this, we used an internal phone list of around 4,000 Enron employees from the Enron Wholesale Group. The phone list contains the name, email address and broad title for these individuals. Since we, again, only have observability

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

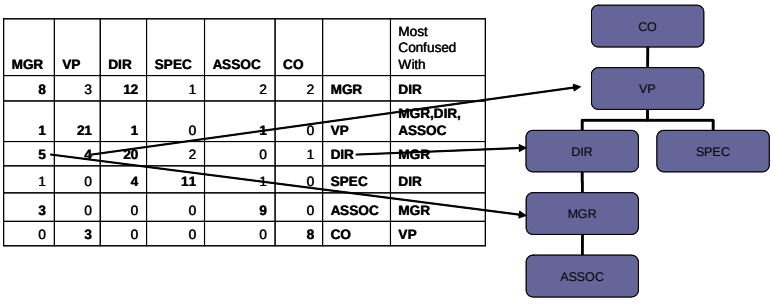

Figure 4. Confusion matrix for broad titles correspond to title hierarchy.

for the archive base, we use only the traffic statistics to and from the archive base as our attributes. Also, to reduce data sparseness for some of these individuals, we limit our set to only those individuals which have exchanged at least 50 emails with the archive base. We also merged levels of a title, such as Junior Specialist, Specialist, and Senior Specialist, and removed any titles which had less than 20 entities. The resulting set contained 416 instances with 5 titles.

6.1. Results

As shown in Figure 5, the results for this larger set are consistent with our earlier results. Here, using 10 fold cross validation, our best performing algorithm is still Bagging with C4.5. We were able to get an accuracy of 55%, still a significant improvement from the random classifier performance of 25%. We also find that our observations regarding the difference between classification of titles hold, with the best classification performance for the highest ranking titles, VP and Specialist, and the lowest title, Analyst. Moreover, the title of Manager is still proving the most difficult to classify.

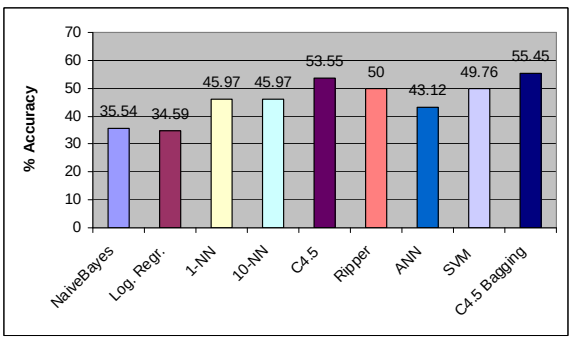

Figure 5. Performance of classifiers on larger EWS employee list.

7. Related Work

7.1. Social Networks

There has been a lot of work in the discovery and analysis of roles and connections between individuals within a social network. In the identification of roles, for example, a significant amount of work has been done to define the centrality and role equivalence of individuals within a social network (Hanneman & Riddle, 2005). More recently, Mccallum et al., presented an LDA approach to topic and role discovery within the network (McCallum & Wang, 2005). Tyler et al, also identified individuals in the head of hierarchy groups using betweeness centrality (Tyler et al., 2005).

With regard to social network connections, McArthur and Bruza used semantic associations to identify implicit and explicit connections between individuals (McArthur & Bruza, 2003). Liben-Nowell and Kleinberg have used co-authorship information to create social networks and predict future interactions within the network (Liben-Nowell & Kleinberg, 2003). Schwartz and Wood generated a social network using the *To* and *From* fields of email messages to discover users of a particular interest and field (Schwartz & Wood, 1992).

7.2. Enron

The release of the Enron data set in 2003 provides researchers a unique opportunity to analyze the traffic of a major corporation. Klimt and Yang provided an overview of this corpus including the number of employees, the number of emails, as well as a generation of a basic social network (Klimt & Yang, 2004b). Moreover, they used the Enron data to explore methods of email classification (Klimt & Yang, 2004a). Corrada-Emmanuel created MD5 hashes of the Enron emails for deduplication and provided mappings

of email account folders in Enron to email addresses (Corrada-Emmanuel, 2004). Using the structure of the emails, Keila and Skillicorn found a relationship in the word use pattern and message length of emails (Skillicorn, 2005). Skillicorn further demonstrated methods to detect unusual and deceptive email communications (Skillicorn, 2005; Fong & Skillicorn, 2006). Diesner and Carley used analysis of the email social network patterns over time to explore crisis detection in email (Diesner & Carley, 2005). More recently, there has been work on the problem of reference resolution by Diehl et al (Diehl et al., 2006). A number of useful tools have also been developed in order to navigate and view email archives (Heer, 2004).

8. Conclusion

In this paper, we discuss the benefits of identifying the hierarchy of the underlying social network 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 underlying social network, using only the structured portion of email communications. We describe a dataset for use in conjunction with the Enron data set for the analysis of the Enron hierarchy. We examine different characterizations of the traffic data information. We then apply various classifiers to this dataset and show that we can get 62% accuracy in classifying individuals to their title, as compared to the 20% accuracy of random. We also show a correlation between the misclassifications and the hierarchical ordering of these titles. These are only our preliminary results, however. We are interested in adding additional attributes for our classification. For one, our experiments only used structured email data, ignoring the content. We would like to use topics and information flow of our content and compare its performance to our results. We also would like to make greater use of commonly used social network analysis measures such as centrality and equivalence in our classifiers. We are also interested in trying to classify more abstract 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

Adibi, J. (2005). Enron dataset. http://www.isi.edu/ adibi/Enron/Enron.htm.

Cohen, W. (2004). Enron email dataset. http://www.cs.cmu.edu/ enron/.

Corrada-Emmanuel, A. (2004). Enron email dataset research. http://ciir.cs.umass.edu/~corrada/enron.

Diehl, C., Getoor, L., & Namata, G. (2006). Name reference resolution in organizational email archives. *6th SIAM Conference on Data Mining*. Bethesda, Maryland.

Diesner, J., & Carley, K. (2005). Exploration of communication networks from the Enron email corpus. *SIAM International Conference on Data Mining*. Newport Beach, CA, USA.

Fiore, A. (2005). Uc berkeley enron email analysis. http://bailando.sims.berkeley.edu/enron_email.html.

Fong, S., & Skillicorn, D. (2006). Detecting word substitution in adversarial communication. *6th SIAM Conference on Data Mining*. Bethesda, Maryland.

Hanneman, R., & Riddle, M. (2005). *Introduction to social network methods*. University of California.

Heer, J. (2004). Exploring Enron: Visualizing ANLP results. http://jheer.org/enron/v1/.

J. Byrne, M. France, W. Z. (2002). At enron, "the environment was ripe for abuse". *Business Week Online*.

Klimt, B., & Yang, Y. (2004a). The Enron corpus: a new dataset for email classification research. *European Conference on Machine Learning/PKDD*. Pisa, Italy.

Klimt, B., & Yang, Y. (2004b). Introducing the Enron corpus. *Conference on Email and Anti-Spam*.

Liben-Nowell, D., & Kleinberg, J. (2003). The link prediction problem for social networks. *Proc. 12th International Conference on Information and Knowledge Management (CIKM)*. New Orleans, Louisiana, USA.

Lyman, P., & Varian, H. (2003). How much information? 2003. http://www.sims.berkeley.edu:8000/research/projects/how-much-info-2003/index.htm.

McArthur, R., & Bruza, P. (2003). Discovery of implicit and explicit connections between people using email utterance. *Proceedings of the Eighth European Conference of Computer-supported Cooperative Work, Helsinki*.

McCallum, A. C.-E. A., & Wang, X. (2005). Topic and role discovery in social networks. *IJCAI*.

Schwartz, M., & Wood, D. (1992). Discovering shared interests among people using graph analysis of global electronic mail traffic. *Communications of the ACM*, *36*, 78–89.

Skillicorn, D. (2005). Detecting unusual and deceptive communication in email. *Centers for Advanced Studies Conference*. Richmond Hill, Ontario, Canada.

Tyler, J., Wilkinson, D., & Huberman, B. (2005). *Email as spectroscopy: Automated discovery of community structure within organizations* (Technical Report).

Witten, I., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*, vol. 2nd Edition. Morgan Kaufmann.