Singapore Management University
Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems School of Information Systems

8-2010

Mining interaction behaviors for email reply order prediction

Byung-Won ON

Ee Peng LIM
Singapore Management University, eplim@smu.edu.sg

Jing JIANG
Singapore Management University, jingjiang@smu.edu.sg

Amruta PURANDARE

Loo Nin TEOW

DOI: https://doi.org/10.1109/ASONAM.2010.60

Citation

ON, Byung-Won; LIM, Ee Peng; JIANG, Jing; PURANDARE, Amruta; and TEOW, Loo Nin. Mining interaction behaviors for email reply order prediction. (2010). *ASONAM 2010: Proceedings: 2010 International Conference on Advances in Social Network Analysis and Mining: 9-11 August 2010, Odense, Denmark*. 306-310. Research Collection School Of Information Systems.
Available at: https://ink.library.smu.edu.sg/sis_research/621

Mining Interaction Behaviors for Email Reply Order Prediction

Byung-Won On, Ee-Peng Lim, Jing Jiang, Amruta Purandare
Singapore Management University
80 Stamford Road, Singapore 178902
{bwon,eplim,jingjiang,amrutadp}@smu.edu.sg

Loo-Nin Teow
DSO National Laboratories
20 Science Park Drive, Singapore 118230
tloonin@dso.org.sg

***Abstract*—In email networks, user behaviors affect the way emails are sent and replied. While knowing these user behaviors can help to create more intelligent email services, there has not been much research into mining these behaviors. In this paper, we investigate user engagingness and responsiveness as two interaction behaviors that give us useful insights into how users email one another. Engaging users are those who can effectively solicit responses from other users. Responsive users are those who are willing to respond to other users. By modeling such behaviors, we are able to mine them and to identify engaging or responsive users. This paper proposes four types of models to quantify engagingness and responsiveness of users. These behaviors can be used as features in the email reply order prediction task which predicts the email reply order given an email pair. Our experiments show that engagingness and responsiveness behavior features are more useful than other non-behavior features in building a classifier for the email reply order prediction task. When combining behavior and non-behavior features, our classifier is also shown to predict the email reply order with good accuracy.**

I. Introduction

A. Motivation

In this paper, we study user interaction behaviors in email networks and how they are relevant to predicting future email activities. An email network is essentially a directed graph with nodes and links representing users and messages from users to other users respectively. Each email is assigned a timestamp and has other attributes including sender, recipients, subject heading, and email content. We focus on two user interaction behaviors that are closely related to how users respond to one another in email networks, namely **engagingness** and **responsiveness**.

We define *engagingness* behavior as the ability of an user to solicit responses from other users, and *responsiveness* behavior as the willingness of an user to respond to other users. A user at the low (or high) extremes of engagingness behavior are known as to be non-engaging (or engaging). Similarly, a user can range from unresponsive to highly responsive. As suggested by their definitions, user engagingness and responsiveness have direct or indirect implications on the way emails are sent and responded, and the strength of relationships users may have with other users in the networks. Nevertheless, these implications have not been well studied. The use of interaction behaviors to enhance email functions has been largely unexplored.

This paper therefore aims to provide a fresh approach towards modeling the engagingness and responsiveness behaviors in email networks. These models are quantitative and assign to each user an engagingness score and a responsiveness score. The scores are within the [0,1] such that 0 and 1 represent the lowest and highest scores respectively. With the scores, we can rank all users by engagingness or responsiveness. Moreover, we derive new features from these behavior scores and use them in an example email activity prediction, e.g., email reply order prediction.

The engagingness and responsiveness behavior models can be very useful in several applications. In the context of business organizations, they help to identify engaging and responsive users who may be good candidates for management roles, and to weed out lethargic users who are neither engaging and responsive making them the bottleneck in the organization. For informal social email networks, engaging and responsive users could be the high network potential candidates for viral marketing applications. Engaging users may solicit more responses for viral messages while responsive users may act fast on these messages. By selecting these users to spread viral messages to targeted user segments by word-of-mouth, marketing objectives can be achieved more effectively.

In this paper, we specifically introduce the **email reply order prediction** task as an application, and show that engagingness and responsiveness behavior models contribute significantly to prediction accuracy. Email reply order prediction refers to deciding which of a pair of emails received by the same user will be replied first. This prediction task effectively helps an email recipient to prioritize his or her replies to emails. For example, if e_1 and e_2 are two emails sent to user u_k who plans to reply both. The outcome of prediction can either be e_1 replied before e_2 or vice versa. The ability to predict reply order of emails has several useful benefits, including helping users to prioritize emails to be replied, and to estimate the amount of time emails get replied. Here, our main purpose is to use the task to evaluate the utility of engagingness and responsiveness behavior models.

We use Enron email dataset in this research. The dataset consists of 517,431 emails from 151 ex-Enron employees. We first preprocess the emails so as to remove noises from the data and to construct the reply and forward relationships among emails. From the email relationships, we also derive email

This work is partially supported by research funding from the DSO National Labs of Singapore.

threads which are hierarchies of emails connected by reply and forward relationships. The email reply order prediction task is addressed as a classification problem. Our approach derives a set of features for a email pair based on the emails' metadata as well as engagingness and responsive behaviors of their senders. As we evaluate the performance of the learnt prediction models, we want to find out the interplay between behavior features and prediction accuracy. Our approach does not depend on email content or domain knowledge which are sometime not available and time costly to process. Given that there are only two possible order outcomes, we expect any method should have an accuracy of at least 50%. In order for email reply order prediction to be useful, a much higher prediction accuracy is required without relying on content analysis.

Both behavior modeling and email reply order prediction are novel problems in email networks. Research on engagingness and responsiveness behaviors is a branch of social network analysis that studies node properties in a network. Unlike traditional social network analysis which focuses on node and network statistics based on static information (e.g., centralities, network diameter) of social networks, behavior analysis is conducted on networks with users dynamically interacting with one another.

In the following, we summarize the important research contributions of this paper:

- We define five behavior models for engagingness and responsiveness behaviors prevalent in email networks. They are (a) email count based, (b) email recipient based, (c) email reply time based, (d) email thread count based, and (e) email reply gap based models. For each model category, one can define different behavior models based on different email attributes. To the best of our knowledge, this is the first time engagingness and responsiveness behavior models are studied systematically.
- We apply our proposed behavior models on the Enron email network, analyze and compare the proposed behavioral models. We conduct data preprocessing on the email data and establish links between emails and their replies. In our empirical study, we found engagingness and responsiveness are distinct from each other. Most engagingness (responsiveness) models of users are shown to be consistent with each other.
- We introduce email reply order prediction as a novel task that uses engagingness, responsiveness and other email features as input features. An SVM classifier is then learnt from the features of training email pairs and applied to test email pairs. According to our experimental results, the accuracy of our SVM classifier is about 69% considering all features except the email sent time. The behavior features alone can achieve 67% accuracy. This indicates that user behaviors are useful in the prediction task.

II. Related Work

There is little research on email user behavior modeling, particularly for engagingness and responsiveness. Prior to this paper, there is only one known work on responsiveness only [1] that defines responsiveness as the deviation in response time of a user from the other users for emails of the same subject. Users with positive and negative deviations are known to be lethargic and responsive respectively.

In the context of prediction tasks for email data, Rowe et al. presented an automatic method for extracting social hierarchy data from the email statistics and structure properties (e.g., degree centrality, number of cliques, etc.) of the email networks [9]. Pathak et al. investigated an socio-cognitive approach to discover knowledge held by users in an email network [8]. This approach involves analyzing who knows who knows who in the network. Karagiannis and Vojnovic studied the prediction of an email reply and response time for a given email represented by features including email size, number of recipients per email, role of the sender and recipient in the organization, information load on the user, etc. [5]. Dredze et al. also proposed a logistic regression model to predict email reply using a variety of email features e.g., dates and times, salutations, questions, and other email header fields [2]. In [11], a supervised classifier was built to automatically label emails with priority levels on the scale of 1 to 5. The features used include graph-based metrics such as node degree, centrality, clique count, and others derived from the underlying social networks of users. McCallum et al. presented the author-recipient-topic model which learns topic distributions based on the direction sensitive messages sent between users [7].

Unlike most previous research on behavior analysis in email networks which focuses on mainly direct statistics of emails such as recipient list size, rate of emails from receiver to sender, and email size to characterize an email user [5], [10], our modeling of engagingness and responsiveness behaviors relies mainly on email reply and forward relationships not available directly in the email data, compared with previous research on email prediction tasks include the prediction of (a) social hierarchy of email users [9], (b) topics of emails [7], and (c) viral emails [10]. Email reply order prediction is thus a new task to be investigated. Although engagingness and responsiveness behaviors and reply order prediction task are defined in the context of email networks, our proposed approaches and results are also applicable to other form of information exchange networks such as messaging and blog networks.

III. Engagingness and Responsiveness Behavior Models

In this section, we describe our proposed behavior models for user engagingness and responsiveness using the notations given in Table I.

A. Email Count Model (EC)

The email count model is defined based on the principle that an engaging user should have most of his/her emails replied, while a responsive user should have most of his/her received emails replied. The engagingness and responsiveness formulas are thus defined by:

$$E^{EC}(u_i) = \frac{|RT(u_i)|}{|S(u_i)|} \tag{1}$$

TABLE I
NOTATIONS.

$S(u_i)$	Emails sent by user u_i
$R(u_i)$	Emails received by u_i
$RB(u_i)$	Email replies sent by u_i
$RT(u_i)$	Emails replying to u_i's earlier emails
$TH(u_i)$	Threads started by an email sent by u_i
$r(e)$	Reply to email e
$Sdr(e)$	Sender of email e
$Rcp(e)$	Recipients (in both To and Cc lists) of email e
$t(e)$	Sent time of email e
$E(u_i \rightarrow u_j)$	Emails from u_i to u_j
$E(u_i \leftrightarrow u_j)$	Emails between u_i and u_j
$rt(u_i \rightarrow u_j)$	Avg. response time from u_i to u_j
$rt(u_i \leftrightarrow u_j)$	Avg. response time between u_i and u_j
$RE(u_i \rightarrow u_j)$	Reply emails from u_i to u_j
$RE(u_i \leftrightarrow u_j)$	Reply emails between u_i and u_j
$MaxRcpCnt$	Largest recipient count

$$R^{EC}(u_i) = \frac{|RB(u_i)|}{|R(u_i)|} \tag{2}$$

When $S(u_i)$ and $R(u_i)$ are empty, $E^{EC}(u_i)$ and $R^{EC}(u_i)$ will be assigned a zero value respectively.

B. Email Recipient Model (ER)

The intuition of this model is that an email with many recipients is likely to expect very few replies. Hence, an engaging user is one who gets replies from many recipients of his/her emails while a non-engaging user receives very few or no reply even when his/her emails are sent to many recipients. On the other hand, a responsive user is one who replies emails regardless of the number of recipients in the emails. A non-responsive user is one who does not reply even if the emails are directed to him/her only. The engagingness and responsiveness formulas are thus defined by:

$$E^{ER}(u_i) = \frac{1}{|S(u_i)|} \sum_{e \in S(u_i)} \frac{|\{u_j \in Rcp(e) \wedge r(e) \in RB(u_j)\}|}{|Rcp(e)|} \tag{3}$$

$$R^{ER}(u_i) = \frac{1}{|R(u_i)|} \sum_{\substack{e \in RB(u_i) \;\; s.t. \\ \exists u_j, \exists e'' \in S(u_j), r(e'') = e}} \frac{|Rcp(e)|}{MaxRcpCnt} \tag{4}$$

where *MaxRcpCnt* is the # of recipients found in the email with largest # of recipients.

C. Email Reply Time Model (ET)

The reply time of an email can be an indicator of user engagingness and responsiveness. The email reply time model adopts the principle that engaging users receives the reply emails sooner than non-engaging users, while responsive users reply to the received emails quicker than non-responsive users. Given an email e' which is a reply of email e, $e' = r(e)$, the *reply time* of e', $RT(e') = t(e') - t(e)$. The z-normalized reply time $\hat{RT}(e')$ is defined by $\frac{RT(e') - \overline{RT}}{\sigma_{RT}}$ where $\overline{RT}$ and σ_{RT} are the mean and standard deviation of reply time respectively. Now, we define the engagingness and responsiveness of *ET* model as:

$$E^{ET}(u_i) = \frac{1}{|S(u_i)|} \sum_{e \in S(u_i)} \frac{1}{|Rcp(e)|} \cdot \sum_{\substack{u_j \in Rcp(e), \\ \exists e' \in RB(u_j), e' = r(e)}} f(\hat{RT}(e')) \tag{5}$$

$$R^{ET}(u_i) = \frac{1}{|R(u_i)|} \sum_{e' \in RB(u_i), e \in R(u_i), r(e) = e'} f(\hat{RT}(e')) \tag{6}$$

where

$$f(x) = \frac{e^{-x}}{1 + e^{-x}} \tag{7}$$

The function $f()$ is designed to convert the normalized reply time to the range [0,1] with 0 and 1 representing extreme slow and extreme fast reply times respectively.

D. Email Thread Count Model (TC)

In the email count model, engagingness is measured by emails sent by a sender and sent emails directly replied by some recipient(s). However, direct reply is not the only type of response to an email. Emails may be indirectly replied in email threads due to forwarded emails. For example, a user u_1 advertises a job position by sending an email to a professor who subsequently forwards the email to his student u_3. If u_3 replies to u_1, we say that the original email is replied indirectly in an email thread. Based on email threads, the thread count model includes indirect replies to emails forwarded between users using the principle: the user is highly engaging if he or she receives many of his/her emails replied directly or indirectly by recipients, and is highly responsive if he or she replies or forwards most emails earlier received. In the following, the engagingness and responsiveness of a user u_i are defined as:

$$E^{TC}(u_i) = \frac{1}{|S(u_i)|} \cdot |\{e \in S(u_i) | \exists t \in TH(u_i), \exists e', e \underset{t}{\twoheadrightarrow} e' \wedge u_i \in Rcp(e')\}| \tag{8}$$

$$R^{TC}(u_i) = \frac{1}{|R(u_i)|} \cdot |\{e \in R(u_i) | \exists u_j, e', t \in TH(u_j), e \underset{t}{\twoheadrightarrow} e' \wedge u_j \in Rcp(e')\}| \tag{9}$$

where $e \twoheadrightarrow_t e'$ returns TRUE when e is directly or indirectly connected to e' in the thread t, and FALSE otherwise.

E. Email Reply Gap Model (RG)

Email sequence refers to the sequence of emails sent and received by a user ordered by time. To derive engagingness and responsiveness from email sequences, we consider the principle that an engaging user is expected to have his or her sent emails replied soon after they are received by the email recipients, and an responsive user replies soon after they receive emails. As users may not always stay online, the time taken to reply an email may vary very much. Instead, we consider the number of emails received later than an email e but are replied before e by a user as a proxy of how soon e is replied.

The above principle is thus used to develop the reply gap model. Let seq_i denote the email sequence of user u_i. When an email received by u_i is replied before other email(s) received earlier, the reply of the former is known as an *out-of-order reply*. Formally, for an email e received by u_i, we define the *number of emails received* and *number of out-of-order replies* between e and its reply e' in seq_i, denoted by $n_r(u_i, e)$ and $n_{\overline{o}}(u_i, e)$ respectively, as

$$n_r(u_i,e) = \begin{cases} \text{\# emails received between} & \text{if } \exists e' \in RB(u_i), \\ e \text{ and } e' \text{ in } seq_i, & r(e) = e' \\ -1, & \text{otherwise} \end{cases} \tag{10}$$

$$n_{\overline{o}}(u_i,e) = \begin{cases} \text{\# emails received} & \text{if } \exists e' \in RB(u_i), \\ \text{between } e \text{ and } e' \text{ in } seq_i & r(e) = e' \\ \text{and have been replied,} & \\ -1, & \text{otherwise} \end{cases} \tag{11}$$

The -1 value is assigned to n_r and $n_{\overline{o}}$ when e is not replied at all. The user engagingness and responsiveness of the *RG* model are thus defined as:

$$E^{RG}(u_i) = \frac{\sum_{e \in S(u_i)} \left(\frac{1}{|Rcp(e)|} \sum_{u_j \in Rcp(e)} \left(1 - \frac{n_{\overline{o}}(u_j,e)}{n_r(u_j,e)}\right)\right)}{|S(u_i)|} \tag{12}$$

$$R^{RG}(u_i) = \frac{\sum_{e \in R(u_i)} \left(1 - \frac{n_{\overline{o}}(u_i,e)}{n_r(u_i,e)}\right)}{|R(u_i)|} \tag{13}$$

IV. Email Reply Order Prediction

We now consider the email reply order prediction which has the following setup. Given a pair of emails (e_i, e_j) sent to the same user from users u_i and u_j respectively, we want to determine the order in which the two emails will be replied. Here, we assume that both e_i and e_j require some replies and u_i and u_j are not the same person. The outcome of prediction is either e_i or e_j first.

Our proposed method is to train a Support Vector Machine (SVM) classifier using labeled email pairs, and to apply the trained classifier on unseen email pairs. For each email pair, we can derive features directly from the emails themselves and their senders including the previous emails they have sent and received. There are three types of features used, namely: (a) *comparative email features* ($\mathbb{E}$), (b) *comparative interaction features* ($\mathbb{I}$) and (c) *comparative behavior features* ($\mathbb{B}$).

Table II lists the email features used in our classifier. For each email feature f_k, we derive a corresponding comparative feature f_k^c of an email pair (e_i, e_j) by

$$(e_i, e_j).f_k^c = e_i.f_k - e_j.f_k$$

. For email send time $t(e)$ feature, we further convert the positive and negative comparative feature values to 1 and -1 respectively. Interaction features refer to set of features derived from the sender of the email to the common recipient u_r as shown in Table III. The behavior features refer to the five E^M and five R^M behavior scores of email senders.

TABLE II
Email Features $\mathbb{E}$.

No	Description	No	Description
1	$t(e)$	9	$\|S(Sdr(e))\|$
2	$size(e)$	10	$\|R(Sdr(e))\|$
3	$size(r(e))$ (assuming we can determine the reply)	11	Avg. $\|S(Sdr(e))\|$ per day
		12	Avg. $\|R(Sdr(e))\|$ per day
4	$size(e) + size(r(e))$	13	$\frac{\|RB(Sdr(e))\|}{\|S(Sdr(e))\|}$
5	$Rcp(e)$	14	$\frac{\|RT(Sdr(e))\|}{\|R(Sdr(e))\|}$
6	$indegee(Sdr(e))$ (# users sending emails to $Sdr(e)$)	15	$\frac{\|RT(Sdr(e))\|}{\|S(Sdr(e))\|}$
		16	$\frac{\|RB(Sdr(e))\|}{\|R(Sdr(e))\|}$
7	$outdegee(Sdr(e))$ (# users receiving emails from $Sdr(e)$)	17	Avg response time for emails in $RT(Sdr(e))$
8	$indegree(Sdr(e))+$ $outdegree(Sdr(e))$	18	Avg response time for emails in $RB(Sdr(e))$

TABLE III
Interaction Features $\mathbb{I}$.

No	Description	No	Description
19	$\|E(Sdr(e) \rightarrow u_r)\|$	27	$\frac{\|RE(Sdr(e) \leftrightarrow u_r)\|}{\|E(u_r \leftrightarrow Sdr(e))\|}$
20	$\|E(u_r \rightarrow (Sdr(e))\|$	28	$rt((Sdr(e) \rightarrow u_r)$
21	$\|E((Sdr(e) \leftrightarrow u_r)\|$	29	$rt(u_r \rightarrow (Sdr(e))$
22	$\|RE((Sdr(e) \rightarrow u_r)\|$	30	# threads involving $(Sdr(e)$, u_j as senders/recipients
23	$\|RE(u_r \rightarrow (Sdr(e))\|$		
24	$\|RE((Sdr(e) \leftrightarrow u_r)\|$	31	# threads involving $(Sdr(e)$, u_r as senders
25	$\frac{\|RE((Sdr(e) \rightarrow u_r)\|}{\|E(u_r \rightarrow (Sdr(e))\|}$		
26	$\frac{\|RE(u_r \rightarrow (Sdr(e))\|}{\|E((Sdr(e) \rightarrow u_r)\|}$		

The comparative interaction and behavior features are defined similar to that of email features.

V. Experiments - Analysis and Comparison of Behavior Models

Our comparison of the proposed user activeness and responsiveness models consists of (a) comparison between different user activeness and responsiveness models, and (b) comparison between activeness and responsiveness.

For (a) and (b), we compare by examining the normalized Kendall tau distance [3]. The Kendall tau distance of two ranked list l_1 and l_2, $\tau(l_1, l_2)$ is defined by:

$$\tau(l_1, l_2) = \sum_{(u_i,u_j) \in U \times U} \hat{\tau}_{u_i,u_j}(l_1, l_2) \tag{14}$$

where $\hat{\tau}_{u_i,u_j}(l_1, l_2) = 0$ if u_i and u_j are in the same order in l_1 and l_2 and 1 if the order is reversed. After normalization by the maximum possible distance, τ value falls between 0 and 1 representing perfect correlation and non-correlation respectively.

Correlation between Engagingness and Responsiveness. We first show the correlation between engagingness and responsiveness for each proposed model using the Kendall τ distance. The τ distance of *EC*, *ER*, *ET*, *TC* and *RG* is 0.46, 0.52, 0.49, 0.46 and 0.5 respectively. These results indicate that engagingness and responsiveness are fairly distinctive behaviors. Most users would receive different ranks for engagingness and responsiveness.

Correlation between different models. Table IV shows the correlations of pairs of models by engagingness and responsiveness respectively. The different engagingness models are quite similar, especially email count model (*EC*) and email thread count model (*TC*). This is due to most email threads

TABLE IV
KENDALL τ DISTANCE BETWEEN ENGAGINGNESS (RESPONSIVENESS) MODELS.

	E^{ER}	E^{ET}	E^{TC}	E^{RG}
E^{EC}	0.14	0.16	0.01	0.18
E^{ER}		0.12	0.14	0.15
E^{ET}			0.16	0.15
E^{TC}				0.18
	R^{ER}	R^{ET}	R^{TC}	R^{RG}
R^{EC}	0.06	0.03	0.01	0.03
R^{ER}		0.07	0.06	0.08
R^{ET}			0.03	0.03
R^{TC}				0.03

TABLE V
RESULTS OF EMAIL REPLY ORDER PREDICTION.

Features used in SVM	Average Accuracy (%)
SVM$_{\mathbb{E}+\mathbb{I}}$	76.68
SVM$_{\mathbb{U}}$	77.04
SVM$_{\mathbb{B}}$	65.67
SVM$'_{\mathbb{E}+\mathbb{I}}$	65.33
SVM$'_{\mathbb{U}}$	68.97

TABLE VI
TOP-10 FEATURES FOR SVM$'_{\mathbb{U}}$.

Rank	Feature	Weight
1	$E^{ET}(Sdr(e_i)) - E^{ET}(Sdr(e_j))$	0.67
2	$R^{RG}(Sdr(e_i)) - R^{RG}(Sdr(e_j))$	0.65
3	$Indegree(Sdr(e_i)) - Indegree(Sdr(e_j))$	0.55
4	$E^{ER}(Sdr(e_i)) - E^{ER}(Sdr(e_j))$	0.49
5	$R^{TC}(Sdr(e_i)) - R^{TC}(Sdr(e_j))$	0.4
6	$\|R(Sdr(e_i))\| - \|R(Sdr(e_j))\|$	0.33
7	$\|E(Sdr(e_i) \rightarrow u_r) - E(Sdr(e_j) \rightarrow u_r)$	0.31
8	$Outdegree(Sdr(e_i)) - Outdegree(Sdr(e_j))$	0.31
9	$size(r(e_i)) - size(r(e_j))$	0.29
10	$size(e_i) - size(e_j)$	0.26

having two to three emails each. The similarity across different models is even more prominent for responsiveness. Again, the *EC* and *TC* models show high correlation in the responsiveness ranking. In particular, our proposed models are correlated by responsiveness rather than by engagingness.

VI. EXPERIMENTS - EMAIL REPLY ORDER PREDICTION ACCURACY

The goal of this experiment is to evaluate the performance our proposed classification approach to predict email reply order. We also want to examine the usefulness of engagingness and responsiveness behaviors in prediction task. There are five SVM classifiers trained, namely: (a) using comparative email and interactive features (denoted by SVM$_{\mathbb{E}+\mathbb{I}}$); (b) using comparative behavior features only (denoted by SVM$_{\mathbb{B}}$), (c) using all features (denoted by SVM$_{\mathbb{U}}$), (d) using comparative email and interactive features except $t(e)$ (denoted by SVM$'_{\mathbb{E}+\mathbb{I}}$), and (e) using all features except $t(e)$ (denoted by SVM$'_{\mathbb{U}}$). Classifiers (d) and (e) are included as earlier study has shown that email replies often follow the last-in-first-out principle. SVM$'_{\mathbb{E}+\mathbb{I}}$ and SVM$'_{\mathbb{U}}$ allow us to find out if we can predict without knowing the email time information. From the 27,730 email reply relationships, we extracted a total of 19,167 email pairs for the prediction task. The emails in each pair have replies that comes after the two emails are received by the same user. For each email pair, we computed feature values based on only email data occurred before the pair. In addition, we used complement email pairs in training. The complement of an email pair (e_i,e_j) with class label c is another email pair (e_j,e_i) with class label $\bar{c}$. Five folds cross validation was used to measure the average accuracy of the classifiers over the five folds. The accuracy measure is defined by $\frac{\#\ correctly\ classified\ pairs}{\#\ email\ pairs}$.

Figure V illustrates the results of all the five SVM classifiers. SVM$_{\mathbb{U}}$ produces the highest accuracy of 77.04% due to the use of all available features. By excluding the email arrival order feature, the accuracy (of SVM$'_{\mathbb{U}}$) reduces to 68.97%. This performance is reasonably good given that random prediction gives an accuracy of 50%. The above results show that email arrival order feature is an important feature in the prediction task. We however notice that behavior features contribute to prediction accuracy especially when the email arrival order feature is not available. Table VI depicts the top 10 features for the SVM$_{\mathbb{U}}$ classifier. The table shows that engagingness based on the email reply time model *ET* is the most discriminative feature. This suggests that engagingness and responsiveness are useful in predicting email reply order.

VII. CONCLUSION

In this paper, we formulate the user engagingness and responsiveness behaviors in an email network. We have developed five behavior models based on different principles. Using the Enron data set, we evaluate these models. We also apply the models to email reply order prediction task and demonstrate that behavior features can be useful in this task. The work is a significant step beyond the usual node and network statistics to determine user behaviors from their interactions. While our results are promising, there are still much room for further research. Enron email dataset is known to have missing emails. We plan to conduct a more comprehensive study on a much larger and complete data set.

REFERENCES

[1] P. Deepak, D. Garg, and V. Varshney. *Analysis of Enron Email Threads and Quantification of Employee Responsiveness.* 2007.

[2] M. Dredze, J. Blitzer, and F. Pereira. Reply Expectation Prediction for Email Management. In *2nd CEAS*, 2005.

[3] R. Fagin, R. Kumar, and D. Sivakumar. Comparing Top K Lists. *SIAM J. on Discrete Mathematics*, 17:134–160, 2003.

[4] T. Karagiannis and M. Vojnovic. Behavioral Profiles for Advanced Email Features. In *WWW*, 2009.

[5] A. McCallum, X. Wang, and A. Coorada-Emmanuel. Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. *Artificial Intelligence Research (JAIR)*, 30(1), 2007.

[6] Nishith Pathak, Sandeep Mane, and Jaideep Srivastava. Who thinks who knows who? socio-cognitive analysis of email networks. In *ICDM*, 2006.

[7] R. Rowe, G. Creamer, S. Hershkop, and S. Stolfo. *Automated Social Hierarchy Detection through Email Network Analysis.* 2007.

[8] S. J. Stolfo, S. Hershkop, C.-W. Hu, O. Nimeskern, and K. Wang. Behavior-Based Modeling and Its Application to Email Analysis. *ACM Trans. on Internet Technology*, 16(2):187–221, 2006.

[9] S. Yoo, Y. Yang, F. Lin, and I. Moon. Mining Social Networks for Personalized Email Prioritization. In *KDD*, 2009.