Evaluating User Actions as a Proxy for Email Significance

Tarfah Alrashed
MIT, USA
tarfah@mit.edu

Chia-Jung Lee
Microsoft, USA
cjlee@microsoft.com

Peter Bailey
Microsoft, Australia
pbailey@microsoft.com

Christopher Lin
Microsoft, USA
christol@microsoft.com

Milad Shokouhi
Microsoft, USA
milads@microsoft.com

Susan Dumais
Microsoft, USA
sdumais@microsoft.com

ABSTRACT

Email remains a critical channel for communicating information in both personal and work accounts. The number of emails people receive every day can be overwhelming, which in turn creates challenges for efficient information management and consumption. Having a good estimate of the significance of emails forms the foundation for many downstream tasks (e.g. email prioritization); but determining significance at scale is expensive and challenging.

In this work, we hypothesize that the cumulative set of actions on any individual email can be considered as a proxy for the perceived significance of that email. We propose two approaches to summarize observed actions on emails, which we then evaluate against the perceived significance. The first approach is a fixed-form utility function parameterized on a set of weights, and we study the impact of different weight assignment strategies. In the second approach, we build machine learning models to capture users' significance directly based on the observed actions. For evaluation, we collect human judgments on email significance for both personal and work emails. Our analysis suggests that there is a positive correlation between actions and significance of emails and that actions performed on personal and work emails are different. We also find that the degree of correlation varies across people, which may reflect the individualized nature of email activity patterns or significance. Subsequently, we develop an example of real-time email significance prediction by using action summaries as implicit feedback at scale. Evaluation results suggest that the resulting significance predictions have positive agreement with human assessments, albeit not at statistically strong levels. We speculate that we may require personalized significance prediction to improve agreement levels.

CCS CONCEPTS

• **Information systems** → **Email**; *Clustering and classification*; *Task models*; • **Human-centered computing** → *User models*;

KEYWORDS

Email communication significance; log data; user activity modelling

WWW 2019, May 13-17, 2019, San Francisco, USA

ACM ISBN 978-1-4503-6674-8/19/05.

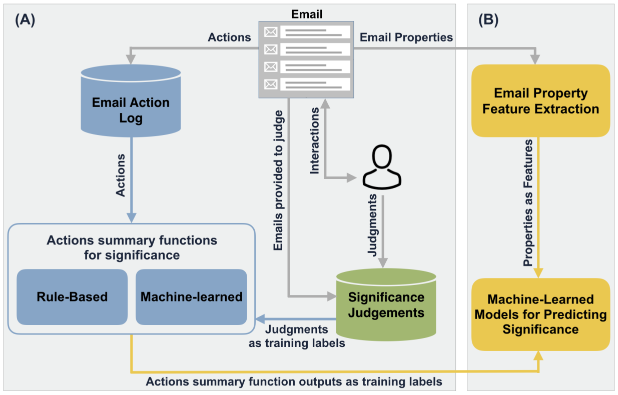
(A)
Email
Actions
Email Properties
(B)
Email Action Log
Actions
Interactions
Emails provided to judge
Judgments
Email Property Feature Extraction
Properties as Features
Actions summary functions for significance
Rule-Based
Machine-learned
Significance Judgements
Judgments as training labels
Machine-Learned Models for Predicting Significance
Actions summary function outputs as training labels

Figure 1: (A) Our framework of evaluating actions as a proxy for email significance, based on action logs and human annotation data. (B) Real-time email significance prediction model trained from actions summary function outputs (§ 7).

ACM Reference Format:
Tarfah Alrashed, Chia-Jung Lee, Peter Bailey, Christopher Lin, Milad Shokouhi, and Susan Dumais. 2019. Evaluating User Actions as a Proxy for Email Significance. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, Article 4, 11 pages.

1 INTRODUCTION

It is a truism that email is still one of the most important means of online communication. While individual email volumes vary greatly, we have observed that work email accounts receive more than 100 emails per day on average, while personal email accounts receive an order of magnitude less on average per day, based on a one-week sample of mailboxes from Microsoft Outlook.[1] To assist people with this amount of incoming email, previous work constructs predictive models to direct attention to emails of potentially higher importance or urgency (e.g. [1, 8, 10, 32]). These models tend to focus on predicting whether some set of strong actions such as Reply or Forward will take place. While such predictors have demonstrated value, strong actions such as Reply comprise only a small subset of the possible sets of actions people can take within their email applications [8]. This may create a gap for identifying important emails because strong actions are not the sole actions indicative of importance (e.g. reading an email several times can represent high importance).

[1] These statistics align with those from a market analysis report by Radicati [22].

In this work, we set out to understand what, if any, relationship exists between the actions carried out on email by people and their perceptions of the significance of those emails. By *significance*, we mean the quality of being worthy of attention that a person accords to an email. Emails that are important, or urgent, or both, are likely to be significant. Other factors may also contribute, such as who it is from or the topic of the email. Our initial analysis based on a large email log sample in Section 3 shows that the actions people perform differ for emails identified as significant versus insignificant. We draw inspiration from an arc of research in the web search community on user actions and inferred document relevance, proceeding from simple clicks [16], to clicks and dwell-time [12], to an entire set of user actions [2].

We conjecture that significance is a quality that exists on a continuous spectrum, and that people make explicit and implicit choices when dealing with their email that reflects this awareness. We also hypothesize that actions carry different semantics and contribute to significance to varying degrees; therefore, diverse actions should be considered when establishing their relationship with significance. Two approaches are investigated for summarizing observed actions in a way that the summaries can then approximate email significance. The first approach presents the notion of weighted action utility (WAU), which is a rule-based fixed-form utility function that summarizes the set of observed actions using a linear combination parameterized on a set of action-specific weights. The second approach attempts to model significance directly using machine learning (ML) techniques.

To evaluate the approaches, we create a human intelligence task (HIT) survey to collect significance judgments from people, including both personal and work accounts. Our results suggest that WAU makes better summaries for determining email significance compared to a reply-only baseline, which is widely adopted as a standard notion for approximating email importance [32] in the literature. As expected, using ML techniques is more effective to make inference over actions and predict email significance. Our follow-up analysis demonstrates that significance prediction using actions varies across users, which may reflect the personalized nature of email activity patterns and significance.

Overall our evaluation results demonstrate that user actions, be it rule-based or machine-learned, can be used as a proxy for users' perceived significance over their own emails. This finding lays an important foundation for downstream tasks such as email prioritization. Building an effective email prioritization predictor with ML usually requires a large quantity of training data, especially given that past research (e.g. [7, 31]) has indicated that email processing is highly personal. Therefore having a means to create unsupervised or semi-supervised training labels at scale is more tractable and practically feasible than aiming for huge quantities of hand-annotated data or explicit user feedback. With the relationship between actions and significance established, we can train such a predictor by leveraging action summaries as a training label, which can be obtained at a low cost by mining action logs. Indeed, in the context of web search, similar ideas of leveraging users' implicit feedback have been studied extensively (e.g. [2, 12, 14, 16]). Specifically, users' actions such as clicks and dwell-time on documents in response to queries have been shown as effective implicit labels of document relevance, based on which standard learning-to-rank techniques are then applicable. To demonstrate this use case, we show an example of how to create a real-time email significance predictor on large-scale email action samples by using the outputs of an action summary function as labels and email properties as features. Our results suggest that the email significance predictions have positive agreement with human assessments, albeit not at statistically strong levels.

To summarize our main contributions, we first evaluate and show that user actions can be used as a proxy for email significance. We consider and conduct extensive experiments on both rule-based and machine-learned techniques to summarize user actions. The results suggest that both approaches outperform a standard reply-only baseline by showing a higher correlation with significance. Figure 1 (A) depicts a high level overview of the process. Furthermore, we identify the analogy between using more than just clicks as implicit feedback for relevance labels in web search and using action summaries as proxies for significance in email processing. In particular, we demonstrate an example of how to leverage an action summary as supervision to train a large-scale email significance predictor, as shown in Figure 1 (B). The predictor can be executed in real-time as it relies on a set of email property features that can be extracted upon email arrival.

2 RELATED WORK

As the volume of email grows and the demand for human attention increases correspondingly, challenges related to email management and retrieval increase [10]. Previous research investigates the different ways that people use and manage their emails in both work and personal settings [6, 11, 31]. As the stream of emails increases, managing email flow is another challenge that has been addressed in the literature [15, 28]. Previous work suggests that with the incoming email flow, people would choose to visit some email messages before others, and spend time selecting which messages to check first [3, 28, 29]. A study has shown that people scan their inbox a couple of times on average before selecting a message to read [4]. Venolia et al. [29] described five stages of email workflow: email flow, triage, task management, archive, and retrieve. Siu et al. [28] extended the work of [29] and their results suggested that people interleave flow, triage and task management, and that handling incoming email involves three steps: glance, scan, and defer. People will usually glance at their inbox several times until a number of new emails have arrived or a message they are expecting has arrived. They would then scan the headers of the new email messages to decide which of these emails need to be read or acted upon. Then they will begin acting on some of these emails messages and/or defer others to return to them later. Although Siu et al. [28] describe the overall steps people take to deal with and manage email flow, it does not address the features people use during the message selecting process and does not describe why people pay more attention to certain email messages over others. Through a user survey and log analysis, Sarrafzadeh et al. [25] investigated how people choose to defer their email for later processing, and whether the decision to defer can be predicted.

Previous work on email prioritization has tried to facilitate the message selection process by including a priority field in messages, helping people pay attention to important emails easily [21]. Other work focused on allowing the sender to assign a priority to a message by adding a "price" to some messages and not others [19, 23]. However, such approaches were not widely adopted, and the priority field was usually ignored by the users. An analysis to understand why people pay attention to some emails over others is needed to better support the development of such systems. Research by Wainer et al. [30] tried to identify why people attend to some emails and not others based on inbox-level cues about message content. In their think-aloud study, they found that individuals make inferences about message content based on top-level cues and that inferred utility as well as curiosity seem to drive attention to a message. In a controlled laboratory experiment they conducted, in which they investigated the relationship between information gap, utility and demand, they found that curiosity drives attention to email under conditions of low demand, and independent of the marked importance of a message. The work by Wainer et al. is an interesting resource to understand what attracts people's attention to their emails when they arrive; however, more insights are needed to be able to predict such attention. Actions on the other hand, have been used as an indicator in several previous works. Machine learning algorithms to predict the likelihood that a message requires a response are described by Yang et al. [32]. That said, the work by Dabbish et al. [8] found that the need to respond is only one part of defining the importance of an email message, and that people responded to information requests or social messages, even though these messages were unimportant for work.

Other email prioritization work focuses on making a personalized prediction of the significance label of emails [33, 34]. Aberdeen et al. [1] used linear logistic regression models to rank mail in the Gmail mail service by how likely the user was to act on an email without explicit labeling from the user. In that model, they used four categories of features: social, content, thread and label features. Neustaedter et al. [21] described a prototype email client that aggregates social meta-data about email correspondents to support email triage. They define metrics for measuring the social importance of users based on elements like recipients' information and email activities, which can be used for predicting relative email ordering.

The work we are presenting here is similar to this line of research in that we are also interested in understanding email prioritization. However, in our work, we are focused just on whether actions that people take on their emails are related to how they perceive significance, rather than attempting to directly introduce a prioritized email system. Our work follows a similar approach to recent work on email search by Kim et al. [18], in which they mapped explicit *in situ* judgments of email search success and effort to implicit actions that people perform on email; instead, we are mapping our users' judgments to email significance. It is also similar to previous work on web search, in which Huang et al. [14] examined mouse cursor behavior on search engine results pages (clicks, cursor movements and hovers over different page regions), as a proxy for relevance; we are using email interactions, specifically actions performed on emails, as a proxy for email significance.

3 ACTIONS AND IMPLICIT SIGNIFICANCE

In an effort to help users be more productive with their email, popular email services have introduced a number of user experience controls to indicate the significance of individual emails. For example, Gmail's Priority Inbox and Outlook's Focused Inbox both attempt to automatically group emails into two categories of importance. These experiences can often provide interesting insights into users' perceptions of email significance, since they allow users to explicitly move emails from one group to another. In the case of Outlook, users can move emails between the Focused and Other tabs. When a user moves an email to the Focused tab, such explicit interaction can say something about the intention of the user – the current email is likely to be more significant. The reversed operation (i.e., moving an email to the Other tab) may implicitly imply that the user considers this email is less significant. We use these explicit interactions with emails to conduct a quantitative study to investigate if action distributions resemble or differ in the two classes of emails, which we refer to as *implicit significant* and *implicit insignificant*. This helps test if actions can be a good proxy for significance. While we understand emails grouped based on these explicit interactions may not necessarily correspond to a user's notion of significance, this is a close surrogate signal for us based on existing user experiences.

To compare the action distributions, we analyze two random, anonymized action log samples provided by Outlook, one for personal and the other for work email activities. In the personal email sample, we have approximately 500 million users and 17 billion messages, and for work emails, we have 170 million users and 8 billion messages. The email service can be accessed from a number of clients, including native apps for both desktop and mobile as well as browser-based interfaces. The logs did not provide access to any content of the email message, email headers, or email search queries. The logs did contain records of the actions performed against the emails, with corresponding timestamps and other metadata such as the client interface type.

Based on these two samples, we analyze the distribution of actions for these two classes, where we consider diverse actions including but not limited to Read, Reply, Forward, Open an Attachment, Click a Link (in the email), Delete, etc. Figure 2 compares the results. The plot suggests that action distributions for implicit significant and implicit insignificant emails are clearly different. This implies that people interact very differently with emails of different properties, as reflected by the corresponding actions. Particular actions, such as replying to a work email or opening an attachment in a personal email might be good indicators of the significance of this email. Conversely, actions such as deleting an email might indicate it to be of low importance. This observation, to some extent, shows supportive evidence for our hypothesis that using actions as significance proxy is sensible.

4 ACTIONS AS A PROXY FOR SIGNIFICANCE

Next we present approaches for summarizing email significance based on a rule-based function and on machine-learned predictors.

4.1 Weighted Action Utility

To summarize the utility obtained from a person's attention on an email, given a set of actions performed on that email, we need some quantity that represents the all-up utility derived from those actions. By *utility* we mean that the attention was useful or beneficial (but

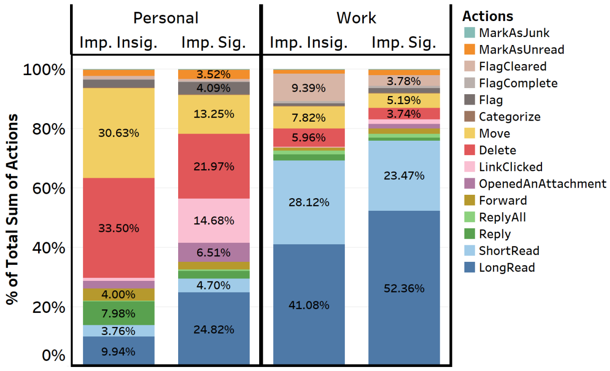

Figure 2: Distribution of actions by implicit insignificant (Imp. Insig.) and implicit significant (Imp. Sig.) emails from a sample of personal and work mailboxes.

not necessarily significant). A naive approach would be to count all actions, giving equal utility for any action. Prior research (e.g. [8]) indicates that different actions (e.g. a Reply to an email vs a Delete) have different degrees of utility. In addition, we can consider that certain compound actions (e.g. a Read followed by a Reply) may indicate that the Read action was more valuable than a Read action in isolation. Thus we also need to allow each type of action (or compound action) to be given a weight.

We propose Weighted Action Utility (WAU), a rule-based fixed-form utility function that summarizes the set of observed actions using a linear combination parameterized on a set of action-specific weights, which is defined as:

$$WAU(e) = \sum_{i=0}^{n} w_{A_i}.A_i \qquad (1)$$

where e is an email, $\{A_1, A_2, ..., A_n\}$ is the set of actions performed on e, and w_{A_i} is the weight associated with an action A_i. There are many ways to determine the weights to be associated with each type of action; here we propose two weight configurations:

4.1.1 Authors' weights. Users can benefit from email in many ways, like by gaining information or organizing their thoughts, and this utility is manifested in the logs through their interactions in their email client (e.g. reading emails, writing emails, pinning emails). Thus, our definition of utility first simply assigns utility scores to a set of user actions and sums these scores over every action taken by the user in their email client. Note however, that some actions do not impart utility by themselves, but in combination with whatever item was acted upon. For example, reading an email that teaches the user something has high utility, whereas reading an email to determine that it is spam provides no utility. Thus, we extend our definition of utility by assigning scores to actions that indicate positive utility if they are preceded by a Read. In other words, whereas normally an action like Pin has zero utility, if it is preceded by a Read, then we assign it some positive utility. The way we assign weights to actions is by determining how important and valuable certain actions are compared to others. For example, we believe that Reply and Reply-all actions are important actions, so we assign them higher weights compared to Delete actions which we do not think should be assigned high value.

4.1.2 Crowd-sourced estimates. People might have very different perspectives from us in terms of how they accord utility to their actions. To obtain a broader and more representative estimate of these utility weights, we ask a pool of 40 people from a private crowd-worker service to assign weights to different types of actions commonly carried out on email. We then take the average of these weights and apply them within WAU. By considering these weights alongside the authors' weights, the objective is to see if different sets of weights can give different results for both or either personal and work emails. Our results and findings are discussed in Section 6.

4.2 Predicting Significance using ML

An alternative method that sidesteps WAU altogether is to learn a model from the actions to the significance annotations created by human directly. One challenge here is whether we can obtain sufficient annotation data to learn a model that does not overfit to the available annotations, given the deeply personal nature of email significance. We will describe the details of the annotation data in Section 5.

We compare a number of ML prediction techniques, including Averaged Perceptron [24], Gradient Boosted Trees [13], and logistic regression. To form a set of useful signals for the model training task, we focus on featurizing actions that have been performed on the emails, including:

- Action unigrams – occurrences of actions in isolation.
- Action bigrams – occurrences of pairs of actions observed for an email. While unigrams reflect the occurrence of each type of action, bigrams may capture specific action sequences that may indicate higher significance of emails (e.g. Flag followed by Reply).
- Time information features including: the total time spent on reading; time to reply; time to read; and time to forward.
- The total number of actions.

We note that we only use action-related features for approximating email significance in this section. As discussed in Section 1, having an action summary function derived purely from user actions can benefit the scalability of large production mail systems. Therefore, features such as the historical interactions between sender and receivers or features that require content analysis, although they could be highly useful, should be investigated separately.

The two methods, WAU and ML, are complementary and together provide a better understanding of how actions can indicate significance. Inferring significance using ML bypasses the need to design a summary function and can potentially result in better predictions when interactions between actions are hidden. However, it is important for production systems to have access to an unsupervised or semi-supervised rule-based "labeler" additionally to create training labels for downstream tasks, and this is where WAU can come into play. Ultimately, the performance of supervised ML predictors will be determined by the scale of data available for training, which for human-annotated emails is generally going to be small-scale and expensive to obtain. Although WAU still poses a few parameters, the simplicity of the formula makes it easy for the weights to be decided elsewhere or heuristically, and also offers high interpretability. A final remark on WAU and ML methods is that they cannot be applied at the time when emails are delivered because at that time no actions are observed. It is similar to web search in that inferring document relevance for a user can only happen after their dwells and clicks. We demonstrate how to apply an action summary in an online prediction task in Section 7.

5 PEOPLE'S PERCEIVED SIGNIFICANCE

Although log data provides implicit characterization of attention utility, people's judgments of the significance of their emails are more valuable for deriving insight. Moreover, as we believe significance is better studied in a continuous spectrum, the fact that current major email clients support only a coarse dichotomy due to binary UI controls is sub-optimal. We developed a human intelligence task (HIT) survey to address these limitations.

The resulting data from HIT can be used as a reference for evaluating action summaries. In the case of an ML action summary, the judgments can also be used as labels in the training phase.

5.1 Email Significance HIT Survey

The email significance HIT survey is a user-email survey that lets people identify or label the absolute significance of individual emails. Each HIT chooses a random email from the most recent 200 emails delivered within the last two months to the user's mailbox, and shows them a rendering of the email's From, Subject, and Body content. It asks them to rate the email by its significance to them at the time they first read it. At the same time, it asks them to select one or more reasons for their decision. Up to 100 emails can be annotated by each person.

Although we believe significance to be a continuous-valued property, for ease of annotation we used a unipolar 5-point Likert scale design. The significance labels ranged from "Insignificant" to "Extremely significant". While the users were seeing their own email messages, we did not record any content of the email. We only collected their significance labels, reasons, and sufficient identifying metadata to let us match the labeled email with its action log record. In addition, we provided users of the HIT survey with the ability to skip any email they felt uncomfortable providing data on to us. While we understand that this might potentially introduce a selection bias towards particular emails, the number of emails that were rated by the users could reduce this bias if any.

Our task guidelines emphasized that rating significance was a personal decision by the participant. We provided suggestions for how to distinguish between different ratings as follows, though not all conditions might exist to select a rating. An "Extremely significant" email is one that communicates essential or important news, or an important and urgent task at work, that the person needed to give their immediate attention to. A "Very significant" email is one that the person would like to pay attention to within the next hour or that the person would return to on multiple occasions within a week after receiving it. A "Moderately significant" email is one the person would like to pay attention to within the next day from receiving it or that they return to at least once. A "Slightly significant" email is one the person would like to pay attention within a week of receiving that email, possibly read or skim the email once, but is unlikely to revisit it again. An "Insignificant" email is an email that the user usually ignored or did not even wish to read.

In the HIT survey we asked people to provide reasons they gave their ratings. The list of reasons we provided to the HIT survey participants includes: Email was from an important person; an Important topic; Part of a conversation I started; Requesting information from me; Information I needed to act on; Of interest but no immediate action; From an organization I often read; I did not have time to pay attention to it; About something I am uninterested in; I usually ignore this sender; and Other. We provided a free-form text field for Other, in case none of the suggested reasons were appropriate. People could select as many reasons as they liked.

5.2 Actions and Perceived Significance

We distributed the survey to both work and personal email participants. For personal emails, we had 118 participants or "judges" and 5774 hits or "judgments". The process of judge selection was managed in a way where bias was minimized, although inevitably they are required to an Outlook user to begin with. Figure 3 shows the distribution of actions (i.e., the percentage of the total number of actions for emails as was ranked "Insignificant" to "Extremely significant"). The most common actions performed on personal emails in all the significance classes are: Read (long and short) and Open an Attachment. A Short Read is one that is 2 seconds or less, to account for read actions that occur when a person is clicking quickly through a list of email, but not stopping to read the content in depth. Long Reads are all reads longer than Short Reads. Separating Short and Long Reads is only one way that attempts to address the complexity of the reading behavior, which is intrinsically affected by different contextual cues such as email length. Both Short and Long Reads are observed at similar rates across different levels of significance. Nevertheless, there is an increase for Open an Attachment actions with the increase of significance, and an increase of the Delete action with decrease of significance. One interesting thing we notice is that Reply and Reply-all have limited impact on how significant an email is. This outcome confirms the results of Dabbish et al. [8] from a study in which they found that the need to respond is only one part of defining the importance of an email message and that people respond to information requests or social messages, even though these messages were unimportant, compared to the work-related messages.

For work emails, we had 24 judges and 560 hits. Figure 4 shows that the distribution of actions for these emails is quite different than their distribution in personal emails shown in Figure 3, as well as the relationship between these actions and the significance of emails. Unlike personal emails, Reply and Reply-all are important actions for work emails. We see an increase of these actions with the increase of the significance of emails.

6 RESULTS AND DISCUSSION

For both WAU and ML predictors, we conduct predictions in two ways. The first casts significance prediction as a binary classification problem, where the outcome directly classifies emails into positive and negative classes of significance. This task mimics what current solutions popular email services provide today. The second considers predicting the graded significance levels obtained from

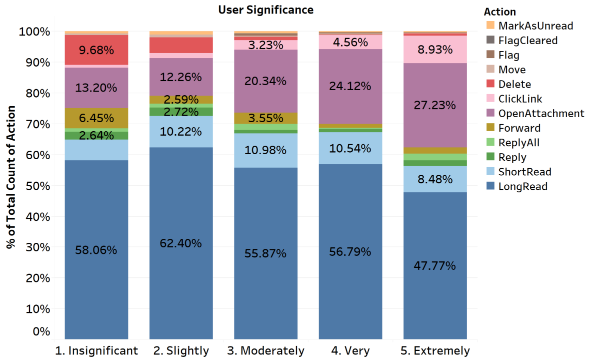

Figure 3: Distribution of actions for personal emails over users' significance

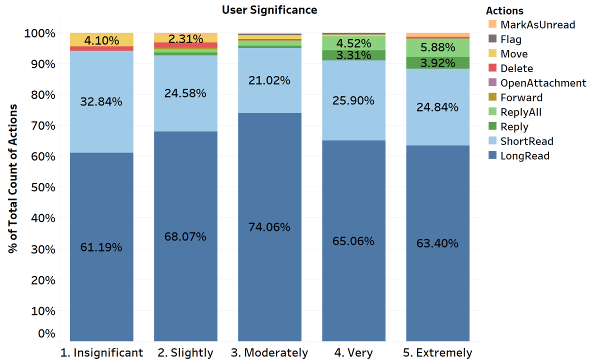

Figure 4: Distribution of actions for work emails over users' significance

the HIT Survey data, based on which a partial ordering or even ranking over the emails can be formed.

6.1 Binary Classification

We organize our five significance labels in two different ways for binary classification:

- Binary Classification Low (BC-Low): positive label includes all significant labels, "Extremely", "Very", "Moderately", and "Slightly" significant. The negative class only includes the "Insignificant" label. This division addresses scenarios where missing out any slightly important email may result in a high penalty.
- Binary Classification High (BC-High): positive class includes "Extremely" and "Very" significant labels, and the negative class includes "Insignificant", "Slightly", and "Moderately" significant labels. This division addresses scenarios where only the top significant emails should draw users' attention.

For our baseline analysis, we classify our emails into emails that have at least one Reply or Reply-all as positive and emails that do not as negative. We choose this reply-only baseline because it is widely adopted as a standard notion for approximating email importance [32]. Table 1 shows the precision, recall and AUC for our baseline analysis for both work and personal emails. Overall the AUC for both email types is around 0.5, which is far from ideal. There is one exception for work emails using BC-High (0.619), suggesting that under certain contexts this reply-only predictor can work well, which is consistent with previous findings. The reason why AUC is higher in BC-High compared to BC-Low is that, looking at Figure 4, emails rated "Very" and "Extremely" significant have more Reply and Reply-all actions than the other emails. One drawback of this simple predictor, however, is that it cannot account for 30% of our judges who have no Reply actions at all on the emails that they judged, even though the emails were in fact assigned with different significance labels.

Table 1: Precision (P), recall (R), and AUC, for our baseline analysis using both BC-Low and BC-High

	Personal			Work		
BC Type	**P**	**R**	**AUC**	**P**	**R**	**AUC**
BC-Low	0.895	0.023	0.509	1.000	0.097	0.548
BC-High	0.416	0.027	0.507	0.625	0.271	0.619

6.1.1 *Using Rule-Based WAU.*

We compute the relationship between the WAU value and the user significance label. We consider the two types of WAU computed using the authors' and crowd-sourced weights. We normalize WAU values using a generalized logistic function. We choose different thresholds (0.1, 0.2, ..., 0.9) to divide the WAU value, where everything beyond the threshold is predicted as positive and everything below is negative. We did not do a train-test split to get an optimal cut because it was important for us to understand the impact of different thresholds.

Table 2 and Table 3 show the prediction results respectively for the tasks BC-Low and BC-High. In both tables, we are only showing 0.4, 0.5, and 0.6 thresholds, due to space limits, and to the fact that extreme thresholds tend to result in more skewed performance (e.g. high precision, low recall). In general, BC-Low appears to be an easier task compared to BC-High (higher precision and AUC values). Among all the thresholds we test, we are getting the highest AUC for both personal and work emails at threshold 0.6. Using BC-Low, the WAU values computed using author's weights can better predict users' significance than the ones computed using crowd-sourced weights in personal emails, and the crowd-sourced weights worked better for work emails. This finding suggests that the same set of weights might not be appropriate to be applied for both types of emails, since personal and work emails tend to have very different action distributions.

6.1.2 *Using Machine-Learned Models.*

We form training and test sets split using 5-fold cross validation on personal and work emails respectively and use McNemar's Test for significance assessment between classifiers. Previous work [33, 34] suggested that personalization is a key factor for improving performance of generically learned models. To study the effects of inter-person variance, when forming the training and test sets, we split the data by either emails or by judges. In both cases, there are no overlapped instances between the splits, and we hypothesize splitting by judges is a more challenging task due to different personal behavior. We experiment with both linear and tree learners, including Averaged Perceptron [24], Boosted Trees [13] and LightGBM [17] for binary

Table 2: Precision (P), recall (R), and AUC, for authors (A) and crowd-sourced (CS) WAU weights(W), using BC-Low with 0.4, 0.5, and 0.6 thresholds (Th)

		Personal			Work		
W	**Th**	**P**	**R**	**AUC**	**P**	**R**	**AUC**
A	**0.4**	0.772	0.991	0.510	0.735	0.995	0.510
	0.5	0.780	0.914	0.530	0.733	0.904	0.502
	0.6	0.905	0.493	0.661	0.863	0.324	0.592
CS	**0.4**	0.621	1.000	0.500	0.731	1.000	0.500
	0.5	0.621	1.000	0.500	0.731	1.000	0.500
	0.6	0.851	0.282	0.601	0.793	0.723	0.610

Table 3: Precision (P), recall (R), and AUC, for authors (A) and crowd-sourced (CS) WAU weights(W), using BC-High with 0.4, 0.5, and 0.6 thresholds (Th)

		Personal			Work		
W	**Th**	**P**	**R**	**AUC**	**P**	**R**	**AUC**
A	**0.4**	0.123	1.000	0.508	0.165	1.000	0.510
	0.5	0.131	0.972	0.541	0.171	0.945	0.530
	0.6	0.212	0.729	0.677	0.240	0.402	0.580
CS	**0.4**	0.248	1.000	0.500	0.164	1.000	0.500
	0.5	0.248	1.000	0.500	0.164	1.000	0.500
	0.6	0.410	0.340	0.590	0.182	0.7391	0.543

Table 4: Precision (P), recall (R), and AUC based on Boosted Trees. Average results are reported with 5-fold cross validation split by emails or by judges, for BC-Low

	Personal			Work		
Split By	**P**	**R**	**AUC**	**P**	**R**	**AUC**
Emails	0.622	0.994	0.643	0.793	0.911	0.698
Judges	0.611	0.983	0.638	0.696	0.900	0.666

Table 5: Precision (P), recall (R), and AUC based on Boosted Trees. Average results are reported with 5-fold cross validation split by emails or by judges, for BC-High

	Personal			Work		
Split By	**P**	**R**	**AUC**	**P**	**R**	**AUC**
Emails	0.603	0.157	0.643	0.517	0.212	0.690
Judges	0.468	0.157	0.626	0.300	0.203	0.600

classification using off-the-shelf ML libraries. In the following analysis, we present results with Boosted Trees only due to space limit.

Table 4 and Table 5 show the prediction results for the tasks BC-Low and BC-High. Compared to Table 2, the performance of boosted trees is in general better than that of fixed cut-off WAU in terms of AUC with *p-value* < 0.01, especially for work emails. This highlights that, while WAU describes a good summary of action distributions, the flexibility of incorporating different dimensions of action metadata, such as action sequences and reading time, can further help significance prediction. The same observation can be found in the task BC-High by comparing Tables 3 and 5.

As expected, cross validation with splits by judges tends to be a harder problem, where AUC decreases between 3% – 15% compared to splits by emails. This observation opens up an interesting research question on how we could use personalization strategies for associating actions and significance, which is beyond the scope of this paper and left for future work.

6.2 Multi-Grade Predictions

Next, we present our findings on prediction tasks where multi-grade significance labels are considered. We examine two prediction strategies. Firstly we form a ranking on emails using the continuous WAU, and compare it with the ranked list based on the five classes of significance collected from the HIT Survey. The second strategy casts the prediction problem as a multi-class classification task, where the outcome aims to distinguish emails into the five significance classes.

6.2.1 Using Rule-Based WAU. We first use the WAU value to order the set of annotated emails. Then we compute Spearman's Rho rank correlation coefficient (with correction for ties) between the ordering formed by WAU and the ordering formed by the five classes of users' significance. Both authors' and crowd-sourced weights are considered. Multi-graded prediction is challenging, thus, as we expected, we get low correlation. For both personal and work emails, we get a correlation that ranges from 0.250 to 0.273 using the two different weights. With binary classification results presented in Tables 2 and 3, we are getting better predictions and correlation with user significance, compared to using the 5-class labels. Nevertheless, computing the correlation per judge (discussed in detail in Section 6.3.2), we observe that some judges have high correlation and others have low correlation, supporting our view that perceptions of significance may be highly personal.

6.2.2 Using Machine-Learned Models. Similar to Section 6.1.2, we conduct 5-fold cross validation and split the data by emails or judges. For the learners, we consider again both linear (multi-class logistic regression) and tree (LightGBM) learners. Micro-accuracy and macro-accuracy are two standard metrics for evaluating multi-class classification effectiveness. Micro-accuracy is defined as the ratio of the number of correctly predicted instances to the total number of instances; this metric is more robust towards class size imbalance problems. Macro-accuracy first computes accuracy for each class and reports the average of per-class accuracies. For our task, we first examine micro-accuracy and macro-accuracy in each fold, take the average of the 5 folds, and report the results in Table 6. Spearman's Rho is additionally computed based on the rankings formed by labels and by predicted classes, so as to capture the correlation between the two. Table 6 shows that predicting graded significance tends to be simpler for personal than work emails. This could be because the action distributions appear more distinguishable across different significance classes in personal than in work data, as suggested by Figures 3 and 4. Compared to the results presented in Section 6.2.1, we see that predicting significance with ML models results in higher correlation with the ground truth rankings for data splits by emails, i.e., Rho=0.315 and Rho=0.295 for personal and work respectively. However, when eliminating personal interactions (i.e., split by judges), the correlation is either

Table 6: Micro-avg and macro-avg accuracy and Spearman's Rho results based on multi-class logistic regression.

	Personal			Work		
Split By	Micro	Macro	Rho	Micro	Macro	Rho
Emails	0.413	0.275	0.315	0.369	0.279	0.295
Judges	0.391	0.245	0.265	0.308	0.241	0.177

on par or worse than using WAU. This outcome may suggest that an evaluation metric which focuses on ranking, i.e. Rho, is more heavily impacted by per-person variance.

6.3 Discussion

While we have demonstrated the potential of summarizing actions to predict email significance, based on both WAU and ML, we observe some challenges which are discussed below.

6.3.1 Differentiating perceived significance. Differentiating perceived significance is challenging in general. As mentioned in Section 5.1, we asked our participants to not only judge and rank the significance of their emails, but to also give us feedback on the reasons for their judgments. While there is usually a clear distinction between "Insignificant" and "Extremely" significant, the distinction between "Slightly", "Moderately", and "Very" significant can be more subtle and highly personal. Asking for participants' feedback was one way to make that distinction clearer for us. Figure 5 shows the distribution of the feedback our participants gave in comparison to their email significance ranking for work emails (personal emails have similar distribution). In both types of emails, the reasons the participants gave for the emails they rated "Extremely" are different than the ones rated as "Insignificant". For "Insignificant" rated emails, we notice that the reasons mostly included: something they are not interested in, they usually ignore the sender, or did not have time to pay attention to it. However, for "Extremely" rated emails, the reasons were typically: an important topic or from an important person, or information they needed to act on. The reasons our participants gave for "Very" and "Extremely" rated emails were similar. This similarity could help explain why we obtain better correlation using the binary classification approach. To some extent, this may also explain why the correlation results presented in Section 6.2 based on correlation using the multi-graded labels are worse in comparison to the binary labels in Section 6.1.

Identifying the same level of perceived significance can be challenging even for the same judge on the same email. A subset of emails were judged repeatedly by the personal judges in the HIT survey. In total 998 unique emails were judged multiple times by 101 judges, with each email being displayed on average 4 times. Note that for all the results presented in Section 6, we remove the duplicates and only experiment with those with a single judgment.

To quantify the degree of inconsistency, we calculate entropy for each unique email based on per-significance distribution using $H(x) = -\sum_{s=1}^{5} P(x_s) log_2 P(x_s)$, where s indicates email significance level[2]. If the outcome is certain (i.e. no inconsistency) we would expect entropy of 0. In fact, we see the average entropy is 0.35 for this data, and gets even higher, i.e., 0.41, when the number of repetitions is greater than or equal to 5. This suggests that email owners may have difficulties giving consistent judgments. Scholer et al. [27] also found high rates of intra-judge search relevance judging inconsistency of between 15% and 24% across a range of TREC test collections.

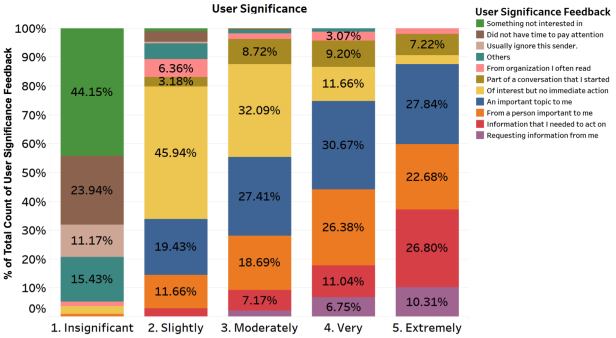

Figure 5: Explanations for the five email significance labels ranked by the HIT Survey participants

6.3.2 Email significance is highly personal. Another challenge we face is that the significance of emails varies based on the user. To better understand the correlation between our WAU significance values and the judged significance, we compute the correlation for each person. Figure 6 shows the correlation for each of the personal email users in our sample data in comparison to the number of hits. We observe that the number of emails the people are judging has no impact on the the correlation. Some people have a high correlation and others have low correlation. More than half of the people have positive correlation, but some of them have no or negative correlation. The large number of zero correlation scores is due to the tie-adjusted Rho computation.

6.3.3 Action Limitations. Looking at our data, both the log data and the HIT Survey labels, we observed two limitations. The same set of actions performed on emails (e.g. emails that have Long Read and Move) can be rated by the HIT Survey participants with different levels of email significance. That is a major challenge given that we depend on actions as features to predict email significance. These set of actions that have different levels of significance vary in terms of both the numbers of this happening and the different labels assigned to them by the judges. We computed the entropy to better understand the variance of each of the actions sets we have and the judges' assigned labels. We noticed that some of the emails that have common action sets (e.g. Long Read, Short Read) have been labeled "Insignificant" to "Extremely" significant by different users, thus have higher entropy (1.5-2). However, the emails that have a unique set of actions were judged using only one label.

7 REAL-TIME SIGNIFICANCE PREDICTION

Above, We studied how and to what extent user actions on emails can be harnessed as a proxy for email significance. We demonstrated that different action summary functions correlate with significance in varying degrees and overall the correlation is positive. To apply our learning to a practical scenario, we train a machine-learned

[2] $0 log 0$ is defined as 0 in information theory.

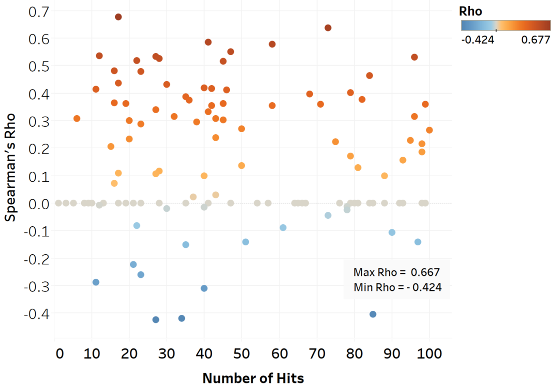

Figure 6: Spearman's Rho for each personal email user over the number of HIT Survey hits

model based on large scale data for real-time significance prediction, as shown in part (B) of Figure 1. Specifically a regression model is trained to fit the output from an action summary function, which is effectively used as the training label. Then we extract a set of non-action features that can be obtained at the time when an email arrives at a mailbox. In this way, the resulting predictive model can forecast email significance in real-time (i.e., email arrival time) because actions are needed only for offline training.

7.1 Prediction Model

Our goal is to have a predictor that is able to predict email significance in real-time. In an ideal world, a supervised predictor can be trained to reach oracle performance if human significance labels can be obtained at scale at a low cost. However, obtaining large amount of human assessments is impractical in practice; instead leveraging implicit feedback from users as a proxy is commonly a preferred, scalable way. In the search world, one of the most successful examples is to treat users' actions including clicks and dwell-time on documents in response to user queries as positive feedback, and thus the corresponding documents are deemed relevant for training [14]. For email significance prediction, we propose to use the output of action summary functions as implicit significance feedback. In this section, we choose to experiment with the rule-based summarizer WAU as the training label, as we have shown that WAU outperforms a reply-only baseline. Although a machine-learned summary function can be more correlated with significance, the amount of human annotation training data heavily determines the performance of the outcome. To focus on presenting the feasibility of building a real-time predictor, a rule-based summarizer is preferred as it provides high simplicity which in turn benefits production systems.

7.1.1 Model training. We reuse the large-scale log samples described in Section 3 to train two real-time predictors respectively for personal and work accounts. In particular, each of the log sample represents an email e_i. We first compute for each e_i its WAU value as the training label y_i based on the actions performed by actual users. Then we extract a number of non-action features $\boldsymbol{x_i}$, resulting in a dataset of $\{(y_i, \boldsymbol{x_i})\}$. We cast the prediction task as a regression problem with an objective that minimizes the $L2$ distance between the prediction $\hat{y_i}$ and y_i. We adopt LightGBM [17] for regression using an off-the-shelf ML library for training. Two held-out datasets (from different users and time periods) in the same format as the samples are added for validation and parameter selection. Parameters on the number of boosting iterations $\{50, 100, 150, 500, 1000\}$, learning rates, and minimum number of items in leaves $\{10, 20, 50\}$ are swept. Training is done on standard CPU machines without parallelization.

7.1.2 Hand crafted email property features. We adopt a total number of 35 non-action features that may reflect email significance. Not all features are available for both personal and work email; we simply impute 0 for missing values. The features are designed in a way that their values can be extracted during email delivery time, which in turn enables real-time significance prediction.

- Sender-related. Binary features that capture if the sender is of high impact, such as if a sender is in the address book or if a sender is more senior in the management chain.
- Recipient-related. Features that capture if the recipient receives an email as a user on the To, CC or BCC field.
- Historical interactions between sender and recipient. Features that reflect in the past how often the sender and recipient interact with each other (e.g., email read rate).
- Binary classifiers. Output from in-house classifiers that predict if the content of an email may belong to a category (i.e., a newsletter, a promotion or a purchase).

We note that any future user action performed on a current email is not considered in this feature list. That is, there is no leak or overlap between the features on email properties and the actual actions that may be performed on this current email at a future time. We should also emphasize that the focus of this section is not to develop a comprehensive set of email property features, but instead to show how online prediction is possible by leveraging output from action summary functions as supervision for ML.

7.2 Evaluation

We compare the real-time prediction $\hat{y_i}$ and the ground truth label $L(e_i)$ on both the personal and work human judgment sets. One way is to compute Spearman's Rho between the rankings formed by predictions and labels as is done in Section 6.2. On the personal set the Rho value is 0.265, and on the work set, it is 0.370. Compared to the results in Section 6.2, we see overall an increase in the correlation for both data sets. One reason for this improvement is that the real-time prediction models tend to create more continuity in the predictions than WAU and the action-features-only models. This continuity breaks ties for emails of similar user responses; the direction of separation usually aligns with the judgments.

In addition to Rho, we also report Krippendorf's α [20] for evaluation on pairwise data formed from the human judgment sets. In the search literature, the amount of quality relevance assessments is key to training an effective ranker. To increase the amount of training data, a common approach is to permute pairs of labelled observations (within some group, such as a query) and incorporate pairwise comparisons in the objective function for optimization (e.g. [5]). In our work, although our goal is not to facilitate learning-to-rank training, we leverage pairwise data generated from the

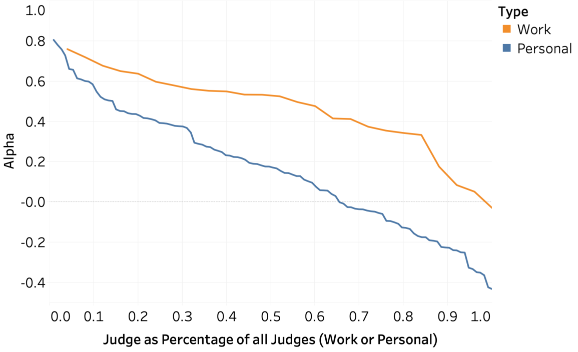

Figure 7: Krippendorf's α for each personal and work email user over their email pairs for significance labels and model prediction scores. The x-axis is the pool size normalized to lie between 0 and 1; agreement rates decline faster with personal judges.

human sets to obtain more observations. Based on that, we can provide evaluation in more detail using Krippendorf's α.

Krippendorf's α is commonly used for assessing agreements between target and reference distributions for various data types (e.g., ordinal, interval and ratio), and deals with ties and missing values naturally. Given a pair of emails (e_i, e_j), $i \neq j$, our goal is to quantify how often the model predictions agree with ground truth ordering. To this end, α is used to assess the degree of consistency between the difference in model scores, $M(e_i) - M(e_j)$, and the difference in human significance labels, $L(e_i) - L(e_j)$. Consistency is higher when the difference in model scores resembles that in labels. For example, when a user regards e_i to be "Very significant" (4) and e_j to be "Insignificant" (1), a model that gives a large positive difference in predictive scores will result in high consistency. The α scores range from -1 (complete anti-agreement) to 1 (complete agreement), where the middle point 0 is effectively random agreement.

We calculate α as the agreement measure on both personal and work judgment data, for all possible pairs of emails belonging to each person. Values for α in aggregate are 0.131 and 0.339 respectively. Overall, we can see that the work data has better agreement in ordering than the personal data, which is consistent with the Rho results. In neither case do the scores show strong levels of agreement; indeed for the personal email corpus it is close to random levels of agreement. (As a comparison, inter-rater agreement for search relevance has also been found to vary considerably, ranging from lows of 0.15 in one study of 180 raters [26] to highs of 0.557–0.705 (depending on condition and scale) in another study involving 16 raters [9], both involving university students.) Search relevance is a challenging labeling task for judges and email significance is also hard for models to predict, based on these outcomes.

In addition to the overall α, we calculate α values on a per person basis. Figure 7 shows these sorted by decreasing α values, and normalized (as a percentage of the pool size) in the x-axis to allow easier comparison between personal and work judges. We see as previously that there is a wide range of agreement according to individuals. Approximately 20% of the work judges and 10% of the personal judges have α values above the 0.667 threshold, which is considered "possibly reliable" for agreement. In the personal case, around a third of the judges start to be anti-correlated. All these findings suggest that, for future work, studying user engagement in depth in personal emails will be important. Indeed, Cecchinato et al [7] found that in personal accounts, people usually looked for order confirmations, travel or money related emails etc. For these cases, it is hard to separate observed user engagement such as reading from other less significant emails such as promotions, given that people also read promotions.

8 CONCLUSIONS

Managing email flow is something that people carry out often. Some emails require more attention than others, and hence, different actions are performed on them accordingly. We investigated our hypothesis that the cumulative set of actions on an email can be considered as a proxy for people's perceived significance of the email using two approaches. The rule-based approach WAU was simple and made better summaries for significance compared to a reply-only baseline; meanwhile the machine-learned approach can more effectively summarize actions and predict email significance.

We found that there are differences in actions between significant and insignificant emails in aggregate. Our analysis also showed that the degree of correlation between actions and significance varies across personal and work accounts as well as across people. Nevertheless, email processing is a much more complex task than traditional web search, in which mapping actions to relevance in search has been shown to have a strong correlation. Thus we carried out more analysis and discussed a number of challenges in our framework, including the fact that the same sets of actions can be observed for emails from different classes of significance. Overall, our results suggest that the cumulative set of actions on any email can be treated as a partial proxy for the email's perceived significance, at least for some people or account types.

Although action summaries are not perfectly correlated with human significance judgments, using actions as a proxy still has potential to be used in predicting significance on emails in real-time, which is similar to modeling document relevance using clicks and dwell time in the context of web search. Specifically, we presented an example of using a rule-based action summary function, WAU, as supervision for training an email significance predictor based on large-scale email action samples, where the resulting predictor can be executed upon email arrival. Evaluation of this predictor showed reasonable correlation with human assessments, more so for work accounts.

Our work provides insights on how the different actions performed on emails can have an impact on the significance of these emails, and how to apply the findings to create a real-time predictor for email prioritization. Future work is required to understand how the personalized nature of significance can be more accurately modeled, including by collecting substantially more annotation data, conducting deeper feature engineering and analysis, or performing *in situ* surveys.

ACKNOWLEDGEMENTS

We gratefully thank Kshitiz Gupta for his help in creating the HIT survey codebase and the anonymous personal and work judges for their judging work.

REFERENCES

[1] D. Aberdeen, O. Pacovsky, and A. Slater. 2010. The learning behind Gmail Priority Inbox. In *LCCC: NIPS 2010 Workshop on Learning on Cores, Clusters and Clouds*.

[2] E. Agichtein, E. Brill, and S. Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proc. SIGIR*. 19–26.

[3] T. Alrashed, A H. Awadallah, and S. Dumais. 2018. The lifetime of email messages: a large-scale analysis of email revisitation. In *Proc. CHIIR*. 120–129.

[4] O. Bälter and C. L. Sidner. 2002. Bifrost Inbox Organizer: Giving users control over the inbox. In *Proceedings of the second Nordic conference on Human-Computer Interaction*. 111–118.

[5] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. 2005. Learning to rank using gradient descent. In *Proc. ICML*. 89–96.

[6] R. Capra, J. Khanova, and S. Ramdeen. 2013. Work and personal e-mail use by university employees: PIM practices across domain boundaries. *JASIST* 64, 5 (2013), 1029–1044.

[7] M. E. Cecchinato, A. Sellen, M. Shokouhi, and G. Smyth. 2016. Finding Email in a Multi-Account, Multi-Device World. In *Proc. CHI*. 1200–1210.

[8] L. A. Dabbish, R. E. Kraut, S. Fussell, and S. Kiesler. 2005. Understanding email use: predicting action on a message. In *Proc. CHI*. 691–700.

[9] T. T. Damessie, F. Scholer, K. Järvelin, and J. S. Culpepper. 2016. The effect of document order and topic difficulty on assessor agreement. In *Proc. ICTIR*. 73–76.

[10] D. Di Castro, Z. Karnin, L. Lewin-Eytan, and Y. Maarek. 2016. You've got mail, and here is what you could do with it!: Analyzing and predicting actions on email messages. In *Proc. WSDM*. 307–316.

[11] N. Ducheneaut and V. Bellotti. 2001. E-mail as habitat: an exploration of embedded personal information management. *Interactions* 8, 5 (2001), 30–38.

[12] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. 2005. Evaluating implicit measures to improve web search. *ACM TOIS* 23, 2 (2005), 147–168.

[13] J. H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* (2001), 1189–1232.

[14] J. Huang, R W. White, and S. Dumais. 2011. No clicks, no problem: using cursor movements to understand and improve search. In *Proc. CHI*. 1225–1234.

[15] K.M. Huessner. 2010. Tech stress: How many emails can you handle a day. Retrieved 14 August, 2018 from https://abcnews.go.com/Technology/tech-stress-emails-handle-day/story?id=11201183

[16] T. Joachims. 2002. Optimizing search engines using clickthrough data. In *Proc. KDD*. 133–142.

[17] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *Proc. NIPS*. 3146–3154.

[18] J Y. Kim, N. Craswell, S. Dumais, F. Radlinski, and F. Liu. 2017. Understanding and Modeling Success in Email Search. In *Proc. SIGIR*. 265–274.

[19] R. E. Kraut, S. Sunder, R. Telang, and J. Morris. 2005. Pricing electronic mail to solve the problem of spam. *Human-Computer Interaction* 20, 1 (2005), 195–223.

[20] K. Krippendorff. 2004. *Content analysis: An introduction to its methodology. Second Edition.* Sage.

[21] C. Neustaedter, A.J. Bernheim Brush, M. A. Smith, and D. Fisher. 2005. The Social Network and Relationship Finder: Social Sorting for Email Triage. In *CEAS*.

[22] S. Radicati. 2014. Email Statistics Report, 2014-2018. Retrieved 14 August, 2018 from http://www.radicati.com/wp/wp-content/uploads/2014/01/Email-Statistics-Report-2014-2018-Executive-Summary.pdf

[23] B. Reeves, S. Roy, B. Gorman, and T. Morley. 2008. A marketplace for attention: Responses to a synthetic currency used to signal information importance in e-mail. *First Monday* 13, 5 (2008).

[24] Frank Rosenblatt. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* 65, 6 (1958), 386.

[25] B. Sarrafzadeh, A.H. Awadallah, C. Lin, C-J. Lee, and M Shokouhi. 2019. Characterizing and Predicting Email Deferral Behavior. In *Proc. WSDM*.

[26] P. Schaer. 2012. Better than their reputation? on the reliability of relevance assessments with students. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 124–135.

[27] F. Scholer, A. Turpin, and M. Sanderson. 2011. Quantifying test collection quality based on the consistency of relevance judgements. In *Proc. SIGIR*. 1063–1072.

[28] N. Siu, L. Iverson, and A. Tang. 2006. Going with the flow: email awareness and task management. In *Proc. CSCW*. 441–450.

[29] G. Venolia, L. Dabbish, J.J. Cadiz, and A. Gupta. 2001. *Supporting email workflow*. Technical Report MSR-TR-2001-88. Microsoft Research.

[30] J. Wainer, L. Dabbish, and R. Kraut. 2011. Should I open this email?: inbox-level cues, curiosity and attention to email. In *Proc. CHI*. 3439–3448.

[31] S. Whittaker and C. Sidner. 1996. Email overload: exploring personal information management of email. In *Proc. CHI*. 276–283.

[32] L. Yang, S. T. Dumais, P. N. Bennett, and A. H. Awadallah. 2017. Characterizing and predicting enterprise email reply behavior. In *Proc. SIGIR*. 235–244.

[33] S. Yoo, Y. Yang, and J. Carbonell. 2011. Modeling personalized email prioritization: classification-based and regression-based approaches. In *Proc. CIKM*. 729–738.

[34] S. Yoo, Y. Yang, F. Lin, and I.-C. Moon. 2009. Mining social networks for personalized email prioritization. In *Proc. KDD*. 967–976.