User Models for Email Activity Management

Mark Dredze
Dept. of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
mdredze@cis.upenn.edu

Hanna M. Wallach
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA, USA
wallach@cs.umass.edu

INTRODUCTION

A single user activity, such as planning a conference trip, typically involves multiple actions. Although these actions may involve several applications, the central point of coordination for any particular activity is usually email. Previous work on email activity management has focused on clustering emails by activity. Dredze *et al.* [3] accomplished this by combining supervised classifiers based on document similarity, authors and recipients, and thread information. In this paper, we take a different approach and present an unsupervised framework for email activity clustering. We use the same information sources as Dredze *et al.*—namely, document similarity, message recipients and authors, and thread information—but combine them to form an unsupervised, non-parametric Bayesian user model. This approach enables email activities to be inferred without any user input. Inferring activities from a user's mailbox adapts the model to that user. We next describe the statistical machinery that forms the basis of our user model, and explain how several email properties may be incorporated into the model. We evaluate this approach using the same data as Dredze *et al.*, showing that our model does well at clustering emails by activity.

DIRICHLET PROCESS CLUSTERING OF EMAILS

Clustering emails by activity involves assigning n email messages $d_1 \ldots d_n$ to k activities $a_1 \ldots a_k$. Each document is represented by a sparse vector, indicating the number of times each word in the vocabulary appears in that document.

One way of modeling these data is to assume that each document was generated by a single activity-specific distribution over words. This model naturally captures the notion that emails about the same activity will use similar words, while emails about different activities will use different words. A Dirichlet process mixture model (DPMM) provides an elegant way of formalizing this idea. Each document d_i is assumed to have been generated by first selecting an activity $a_i = j$ for that document and then drawing words from the activity-specific distribution over words $\theta^{(j)}$. This process may be inverted using statistical inference techniques, allowing the unknown activity assignments a_i and activity-specific distributions over words $\theta^{(j)}$ to be inferred from a set of unlabeled documents. Furthermore, advance specification of the number of activities is not required—this is automatically determined from the data.

A user-specific DPMM may be constructed by clustering the emails in the user's inbox into activities—future emails will be assigned to activities on the basis of this user-specific clustering. The latent activity assignments may be inferred using Gibbs sampling [4] as follows.

The probability of assigning document d_i to activity j is

$$\begin{aligned} &P(a_i = j \,|\, a_{-i}, d_1, \ldots, d_n) \\ &\quad \propto P(a_i = j \,|\, a_{-i}) P(d_i \,|\, d_{-i}, a_i = j, a_{-i}) \end{aligned} \tag{1}$$

where d_{-i} denotes all documents excluding d_i and a_{-i} denotes the activity assignments for these documents.

The first term, $P(a_i = j \,|\, a_{-i})$, is the prior probability of choosing activity j and is given by

$$P(a_i = j \,|\, a_{-i}) = \begin{cases} \frac{N_j}{\alpha + N.} & j \text{ exists} \\ \frac{\alpha}{\alpha + N.} & j \text{ is new,} \end{cases} \tag{2}$$

where N_j is the number of documents assigned to activity j (excluding d_i) and $N. = \sum_j N_j$. The parameter α determines the rate with which new activities are created. *A priori*, document d_i is more likely to be assigned to an activity that already has many documents associated with it.

The second term, $P(d_i \,|\, d_{-i}, a_i = j, a_{-i})$, is the probability that document d_i was generated by activity j, given all other activity assignments. This enforces the requirement that documents containing similar words be assigned to the same activity. $P(d_i \,|\, d_{-i}, a_i = j, a_{-i})$ may be computed by marginalizing over all possible values of $\theta^{(j)}$ under a symmetric Dirichlet prior with scaling parameter β:

$$\begin{aligned} &P(d_i \,|\, d_{-i}, a_i = j, a_{-i}) \\ &\quad = \int d\,\theta^{(j)} P(d_i \,|\, \theta^{(j)}) P(\theta^{(j)} \,|\, d_{-i}, a_{-i}) \\ &\quad = \frac{\Gamma(\beta + N._{\,|\, j})}{\prod_w \Gamma(\frac{\beta}{W} + N_{w \,|\, j})} \frac{\prod_w \Gamma(\frac{\beta}{W} + N_{w \,|\, j} + M_w)}{\Gamma(\beta + N._{\,|\, j} + M.)}, \end{aligned} \tag{3}$$

where $N_{w \,|\, j}$ is the number of times word w has been used in all the documents assigned to activity j (excluding d_i), M_w is the number of times w has been used in d_i, $N._{\,|\, j} = \sum_w N_{w \,|\, j}$ and $M. = \sum_w M_w$.

GENERALIZING THE CLUSTERING PRIOR

Equation 2 is the distribution over activities under a Dirichlet process prior, however, there are other clustering priors that might be more appropriate for clustering email messages by activity. In the most general setting,

$$P(a_i = j \mid a_{-i}, x_i, x_{-i}) \propto \begin{cases} f(N_j, x) & j \text{ exists} \\ \alpha & j \text{ is new,} \end{cases} \tag{4}$$

where x_i is additional data for document d_i, upon which the model is conditioned. For example, the x_i might be the set of authors and recipients of d_i. This general prior reduces to a Dirichlet process prior when $f(N_j, x) = N_j$.

People

The emails within a single activity are typically associated with a small set of people. Any new email that is sent to or from these people (or some of these people) is likely to be part of this activity. Consequently, using an activity clustering prior that incorporates information about message authors and recipients will probably result in better activity clusters than those obtained using a Dirichlet process prior.

Dredze *et al.* [3] introduced SimSubset, a similarity metric for scoring email-activity pairs:

$$s_p(\mathcal{P}_i, j) = \frac{|\mathcal{P}_i \bigcap \left(\bigcup_{d_k \mid a_k = j, k \neq i} \mathcal{P}_k\right)|}{|\mathcal{P}_i|}. \tag{5}$$

$\mathcal{P}_i$ is the set of people associated with document d_i, while $\bigcup_{d_k \mid a_k = j, k \neq i} \mathcal{P}_k$ is the set of people associated with all other documents assigned to activity j. When $\mathcal{P}_i = \emptyset$ define $s_p(\mathcal{P}_i, j) = 1$ When everyone in $\mathcal{P}_i$ is also associated with the set of documents in activity j, $s_p(\mathcal{P}_i, j)$ will be one. When none of the people associated with d_i are associated with activity j, $s_p(\mathcal{P}_i, j)$ will be zero.

The SimSubset metric can be incorporated into an activity clustering prior as follows:

$$\begin{aligned} &P(a_i = j \mid a_{-i}, \mathcal{P}_1, \dots \mathcal{P}_n) \\ &\quad \propto \begin{cases} N_j \, s_p(\mathcal{P}_i, j) & j \text{ exists} \\ \alpha & j \text{ is new.} \end{cases} \end{aligned} \tag{6}$$

When everyone in $\mathcal{P}_i$ is associated with existing activity j, activity j is chosen with probability proportional to N_j. When no one in $\mathcal{P}_i$ is associated with activity j, the probability of choosing activity j is zero.

Threads

Email thread information is another good predictor of activities [3]. Using the same approach as above, the prior over activity clusters can be changed so as to incorporate information about threading. Given a document d_i and a corresponding thread indicator r_i, a thread-dependent metric for scoring email-activity pairs may be defined as follows:

$$s_r(r_i, j) = \begin{cases} 1 & r_i \text{ is a single-mail thread} \\ N_{r_i \mid j} + \kappa & \text{otherwise.} \end{cases} \tag{7}$$

$N_{r_i \mid j}$ is the number of documents belonging to thread r_i that are assigned to activity j (excluding d_i). κ is a parameter that ensures that activities that do not contain any messages from this thread still receive a (small) non-zero score.

This metric can be incorporated into an activity clustering prior as follows:

$$\begin{aligned} &P(a_i = j \mid a_{-i}, r_1, \dots r_n) \\ &\quad \propto \begin{cases} N_j \, s_r(r_i, j) & j \text{ exists} \\ \alpha & j \text{ is new.} \end{cases} \end{aligned} \tag{8}$$

Activity j is chosen with probability proportional to the product of N_j and $s_r(r_i, j)$. When d_i is the only document in its thread, the probability of choosing activity j is proportional to N_j as in the Dirichlet process prior.

TOPICS

Equation 3 arose from the assumption that each document was generated by first choosing an activity $a_i = j$ for that document and then drawing words from the activity-specific distribution over words $\theta^{(j)}$. In fact, other methods of document generation may yield better activity clusters. One such method is latent Dirichlet allocation [2]. Latent Dirichlet allocation (LDA) treats each document as a finite mixture over an underlying set of topics, where each topic is characterized as a distribution over words. For example, an email inbox might contain latent topics that correspond to concepts such as "user modeling," "hotel rooms" and "flights." Each email has a different distribution over these topics: an email about going on vacation might give equal probability to the last two topics, while an email about attending a user modeling workshop might give equal probabilities to all three. LDA assumes that each word in a document d_i is generated by first sampling a topic t from a document-specific distribution over topics $\theta^{(d_i)}$ and then sampling a word from the topic-specific distribution over words $\phi^{(t)}$.

It is possible to incorporate LDA into the user model by assuming that each document d_i was generated by first sampling an activity for that document and then generating each word by drawing a topic from an activity-specific distribution over topics and a word from the corresponding topic-specific distribution over words. Inverting the procedure yields a new probability of assigning document d_i to activity j:

$$\begin{aligned} &P(a_i = j \mid a_{-i}, d_1, \dots, d_n, z_1, \dots, z_n) \\ &\quad \propto P(a_i = j \mid a_{-i}) P(z_i \mid z_{-i}, a) P(d_i \mid d_{-i}, z), \end{aligned} \tag{9}$$

where z_i is the set of topic assignments for document d_i.

The second term, $P(z_i \mid z_{-i}, a)$, is given by

$$\frac{\Gamma(\beta + N_{\cdot \mid j})}{\prod_t \Gamma(\frac{\beta}{T} + N_{t \mid j})} \frac{\prod_t \Gamma(\frac{\beta}{T} + N_{t \mid j} + M_t)}{\Gamma(\beta + N_{\cdot \mid j} + M_{\cdot})}, \tag{10}$$

where $N_{t \mid j}$ is the number of times topic t has been used in all documents assigned to activity j (excluding d_i), M_t is the number of times t has been used in d_i, $N_{\cdot \mid j} = \sum_t N_{t \mid j}$ and $M_{\cdot} = \sum_t M_t$. Similarly, the third term, $P(d_i \mid z)$, is

$$\begin{aligned} &\prod_t \frac{\Gamma(\gamma + N_{\cdot \mid t})}{\prod_w \Gamma(\frac{\gamma}{W} + N_{w \mid t})} \\ &\qquad \times \frac{\prod_w \Gamma(\frac{\gamma}{W} + N_{w \mid t} + M_{w \mid t})}{\Gamma(\gamma + N_{\cdot \mid j} + M_{\cdot \mid t})}, \end{aligned} \tag{11}$$

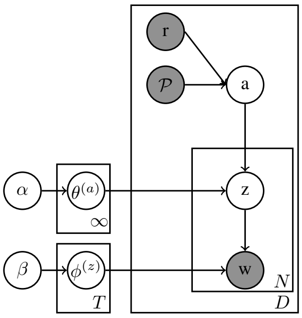

Figure 1. The combined user model. Shaded nodes indicate observed variables. Each document d_i is generated by first sampling an activity a_i given the set of people $\mathcal{P}_i$ and thread indicator r_i associated with that document, and then generating each word by drawing a topic from an activity-specific distribution over topics and a word from the corresponding topic-specific distribution over words.

where $N_{w\,|\,t}$ is the number of times word w has been assigned to topic t in all documents excluding d_i, $M_{w\,|\,t}$ is the number of times w has been assigned to topic t in d_i, $N_{\cdot\,|\,t} = \sum_w N_{w\,|\,t}$ and $M_{\cdot} = \sum_w M_w$.

COMBINED MODEL

It is possible to define a single activity clustering prior that combines thread information with information about message authors and recipients. Under this prior, the probability of choosing activity j is given by

$$P(a_i = j \,|\, a_{-i}, \mathcal{P}_1, \ldots \mathcal{P}_n, r_1 \ldots r_n) \propto \begin{cases} N_j \, s_p(\mathcal{P}_i, j) \, s_r(r_i, j) & j \text{ exists} \\ \alpha & j \text{ is new.} \end{cases} \tag{12}$$

This may be used as the first term in equation 9, resulting in a non-parametric Bayesian model of email activities that incorporates information about threads, people, and topics. The corresponding graphical model is shown in figure 1.

Since topics and activities are both unknown, the activity clusters must be inferred by alternating between sampling activities given topics using equation 9 and sampling sampling topics given activities. The topic assignments may be initialized using LDA. The hyperparameters α and κ are optimized using evidence maximization [5].

EVALUATION

All model variants were evaluated using the same data as Dredze *et al.* [3]. This data set consists of 1175 messages sent over a period of ninety days. Two hundred of the messages are labeled with email activities. In total, there are twenty-seven activities, which range in size from one to thirty-eight messages. Quoted text was removed from most messages. Message subjects and bodies were combined, and stop words and punctuation were removed.

Information about threads, message authors and recipients was collected as follows:

- **People:** The email addresses in each message's "From," "To" and "CC" fields were extracted, creating a set of 1675 unique addresses. The user's address was removed since it was included on almost every email.
- **Thread:** Simple threading, using "Message-Id," "In-Reply-To" and "References" headers, resulted in 823 threads.

Method	*Diver.*	*Disp.*	*Mean*	*Activities*
DP	1.241	1.4695	1.3552	247
People	0.8951	1.5293	1.2122	313
Thread	1.2316	1.4573	1.3445	244
DP+Topics	0.6637	1.6773	1.1705	312
Combined model	0.4330	1.67	1.0515	354

Table 1. Results for activity clustering in diversity, dispersion and their mean (smaller is better). Results averaged over 5 runs.

All model variants were randomly initialized by assigning documents to one of 1000 activities. Unnecessary activities were automatically eliminated during activity inference. Each iteration involved twenty complete Gibbs passes through the data, sampling activities. If a model variant used topics, 200 topic-sampling Gibbs passes were also performed. In total, inference consisted of thirty iterations.

The model variants were compared by computing the diversity and dispersion of the activity clusterings [1]. Diversity measures the extent to which clusters consist of messages from a single real activity—a diversity score of zero indicates that for every cluster, all messages in that cluster belong to the same real activity. Dispersion measures the extent to which a single real activity is dispersed among the clusters—a dispersion score of zero means that for every real activity, all messages from that activity appear in the same cluster. Results for each model variant are shown in table 1, along with the number of activities inferred by each model. The email-specific clustering priors outperformed the Dirichlet process prior. Adding topics also improved performance. Combining all three sources of information—threads, people and topics—substantially increased performance over the other model variants.

ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by DoD contract #HM1582-06-1-2013. Any opinions, findings and conclusions or recommendations expressed in this material are the author's and do not necessarily reflect those of the sponsor.

REFERENCES

1. I. Bhattacharya and L. Getoor. Entity resolution in graph data. Technical Report CS-TR-4758, University of Maryland, College Park, October 2005.
2. D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
3. M. Dredze, T. Lau, and N. Kushmerick. Automatically classifying emails into activities. In *IUI*, 2006.
4. R. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.
5. X. Zhu, Z. Ghahramani, and J. Lafferty. Time-sensitive Dirichlet process mixture models. Technical Report CMU-CALD-05-104, Carnegie Mellon University, 2005.