CC Prediction with Graphical Models

Chris Pal and Andrew McCallum
Dept. of Computer Science
140 Governors Drive
University of Massachusetts
Amherst, MA 01003-9264, USA
{pal, mccallum}@cs.umass.edu

ABSTRACT
We address the problem of suggesting who to add as an additional recipient (i.e. cc, or carbon copy) for an email under composition. We address the problem using graphical models for words in the body and subject line of the email as well as the recipients given so far on the email. The problem of cc prediction is closely related to the problem of expert finding in an organization. We show that graphical models present a variety of solutions to these problems. We present results using naively structured models and introduce a powerful new modeling tool: plated factor graphs.

1. INTRODUCTION
There are many important situations in which people composing email may wish to have an automated system suggest a list of additional recipients to cc. For example, if a user is working on a project they may have forgotten to include a team member, collaborator or manager on an email. In another scenario, an author may wish to identify people within their organization or social network who are working on similar projects, dealing with similar issues or who have relevant skills.

The ability to identify people to cc who are outside of ones normal pattern of email communication also has great potential help organizations avoid “stovepiping”. A stovepipe organization contains members who have narrowly defined responsibilities and information, output and feedback only moves along a set path through a management hierarchy. An organization can potentially be more adaptive when stovepiping is avoided. For these reasons we are interested in constructing a principled system for cc prediction.

2. MODELS AND SYSTEMS
In our work here we begin with a simple multinomial naive Bayes model [5] for words in the body of the message under composition. To train this model for cc prediction, for each email in a users sent mailbox we consider each recipient as a target label and replicate emails where necessary. Figure 1 (Left) illustrates a classic naive Bayes document model involving n draws from discrete random variables $x_i, i = 1 \ldots n$ for each of the words in the document using factor graph notation [3]. Importantly, there are a different number of words n for any possible email. To use this model for prediction we simply instantiate n observed words for an email under composition, use the model to compute the distribution over labels $y$, and present the user with a list sorted by its probability. We can then extend this construction using graphical models to capture the richer structure of email.

Other work has looked at extending the standard naive Bayes model for document classification. For example, in [7] a scoring function was proposed involving different “weights” for the contributions of underlying words arising from the message body and words arising from the message subject. This approach also normalized for message length. In our approach here, we partition the emails into three different sections, the body, the subject and the recipients. We then use three different discrete conditional distributions for variables observed within these different sections. For an email under consideration we thus have $N_b$ and $N_s$ words in the body and subject respectively. Again, for each of the $N_r$ recipients we replicate the email (simulating what actually happens when an email with multiple recipients is sent). Each replication has a different recipient as the target along with the remaining $N_r-1$ recipients. We process email addresses into a bag of words breaking at periods, spaces and @. This gives us some tolerance for minor perturbations of email addresses when identity resolution is inexact. We do not distinguish between the recipients in the TO and CC fields as our previous investigations have found little utility in making the distinction.

We propose illustrating these types of models using a combination of factor graphs and “sheet” or plate notation [10]. Plates are widely used to compactly illustrate replicated variables in Bayesian networks. Plated factor graphs allow mixtures of undirected and directed graphical models to be compactly illustrated. However, for our experiments here we use locally normalized factors. Figure 1 (Right) illustrates our extended model using plated factor graph notation. To the best of our knowledge this is the first presentation of a plated factor graph. Plated factor graphs also have the advantage that the details of function factorizations and replications are more explicitly illustrated in the graph.

3. EVALUATION
We used the personal email data set of McCallum which was also used for experiments presented in [4]. Emails in this set were generated between January 3 to October 10, 2004. There are 825 unique users in the corpus after accounting for multiple email addresses. We use the sent mailbox of this corpus which consists of 9244 messages. The size of the
uncertainty problems under similar situations in a more fully.

Various methods have been proposed to deal with identity
potential to capture some of these address variations.

Our bag of words representation for addresses therefore has
different machines each producing variations of their address.

arises due to the fact that users typically send email from dif-
step in order to obtain models with good performance. This
lar we have found that identity resolution is a very important
messages deeper along the thread.

coding the email address information for recipients within
email is in reply to another email, the third row of Table
increasing cc prediction performance. As well, when a given
addition of co-recipient information was a dominant factor
predictions for each email in the sent mailbox over the course of
each day at 4:00am. We evaluate models by making cc pre-
dictions using the remaining recipients as observations.

We score a cc prediction as correct if the held out recipient is
predictions using the remaining recipients as observations.

We have used factor graphs with locally normalized fac-
tors for our experiment here because parameter estimation
amounts to computing sufficient statistics which can be quickly
computed. This also leads to fast incremental estimation
which is an important design criterion that enables a system
to rapidly adapt as new email is generated. Another advan-
tage of using the plated factor graph notation is that we can
describe models that are not locally normalized as well as
models that are obtained via discriminative optimization as
is done in the Conditional Random Field (CRF) framework
[8]. One can therefore extend the CRF framework to plated
factor graphs. As well, hybrid generative/discriminative
methods such as multi-conditional learning [6] and semi-
supervised methods are straightforward to derive within the
plated factor graph framework. However, fast incremental
training methods that can deal with thousands of potential
output labels and tens of thousands of features are desir-
able for many real world cc prediction scenarios. We found
that standard gradient based optimization methods for the
analogous multinomial logistic regression models defined by
these graphical structures were unacceptably slow but see
potential for future investigation.

4. ANALYSIS AND DISCUSSION

In the experiments shown in Table 1 we found that the
addition of co-recipient information was a dominant factor
increasing cc prediction performance. As well, when a given
email is in reply to another email, the third row of Table
1 illustrates the effect of the addition of a fourth plate en-
coding the email address information for recipients within
the previous email using a bag of words. The effect here
was small and we are presently investigating features from
messages deeper along the thread.

For email exchanges in academic environments in particu-
lar we have found that identity resolution is a very important
step in order to obtain models with good performance. This
arises due to the fact that users typically send email from dif-
ferent machines each producing variations of their address.
Our bag of words representation for addresses therefore has
the potential to capture some of these address variations.
Various methods have been proposed to deal with identity
uncertainty problems under similar situations in a more fully

automated fashion [9]. However, we have integrated the cc
prediction models presented here into the larger CALO sys-
tem [2]. In CALO, a moderate level of identity resolution
or reification is presently performed by other components of
the system. User specified ground truth identity resolution
was therefore important for our experiments here. However,
the raw email addresses that will be used for formal system
tests will likely have less variability than is observed in our
test set here.

We have used factor graphs with locally normalized fac-
tors for our experiment here because parameter estimation
amounts to computing sufficient statistics which can be quickly
computed. This also leads to fast incremental estimation
which is an important design criterion that enables a system
to rapidly adapt as new email is generated. Another advan-
tage of using the plated factor graph notation is that we can
describe models that are not locally normalized as well as
models that are obtained via discriminative optimization as
is done in the Conditional Random Field (CRF) framework
[8]. One can therefore extend the CRF framework to plated
factor graphs. As well, hybrid generative/discriminative
methods such as multi-conditional learning [6] and semi-
supervised methods are straightforward to derive within the
plated factor graph framework. However, fast incremental
training methods that can deal with thousands of potential
output labels and tens of thousands of features are desir-
able for many real world cc prediction scenarios. We found
that standard gradient based optimization methods for the
analogous multinomial logistic regression models defined by
these graphical structures were unacceptably slow but see
potential for future investigation.

5. FUTURE WORK

We are presently evaluating the utility of using informa-
tion from within a users incoming email as well as features
from org charts and various other relationships to increase
cc prediction performance. As well, the framework we have
presented here can be extended in a straightforward manner
to assist with the identification of people within an organiza-
tion or community who could be cc’d but for whom a given
user may have never corresponded with over email. The
modularity of graphical models provides us with a frame-
work for enabling this scenario whereby model parameters
for word usage associated with people can be shared among
users. In another possible scenario we could take the pop-
ular approach of introducing hidden topic variables [1] into
our model or add more detailed sender and recipient struc-
ture as in the construction of [4]. One could then enable
users to select topics consisting of word lists that they are
willing to share with different members of their organization
or community. These distributions could then be integrated
into the graphical model.

Table 1: A comparison cc prediction accuracy for
naive Bayes models and plated factor graph models.

<table>
<thead>
<tr>
<th>Model</th>
<th>First Month</th>
<th>Last Month</th>
<th>Avg. Daily</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>.301</td>
<td>.326</td>
<td>.364</td>
</tr>
<tr>
<td>Factor Graph</td>
<td>.364</td>
<td>.395</td>
<td>.448</td>
</tr>
<tr>
<td>Thread Info</td>
<td>.357</td>
<td>.403</td>
<td>.448</td>
</tr>
</tbody>
</table>
6. ACKNOWLEDGEMENTS

We thank Rich Giuli for discussion about the integration of these methods within the CALO system and other insights. We also thank Adam Cheyer for some initial motivation for investigating the question of cc prediction. This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010 and Microsoft Research under the Memex funding program. Any opinions, findings and conclusions or recommendations expressed in this material are the authors’ and do not necessarily reflect those of the sponsor.

7. REFERENCES


