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 mail box we consider each recipient as a target label y and replicate emails where necessary. Figure 1 (Left) illustrates a classic naive Bayes document model involving n draws from discrete random variables $x_i, i = 1 \dots n$ 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

CEAS 2006 - Third Conference on Email and Anti-Spam, July 27-28, 2006, Mountain View, California USA

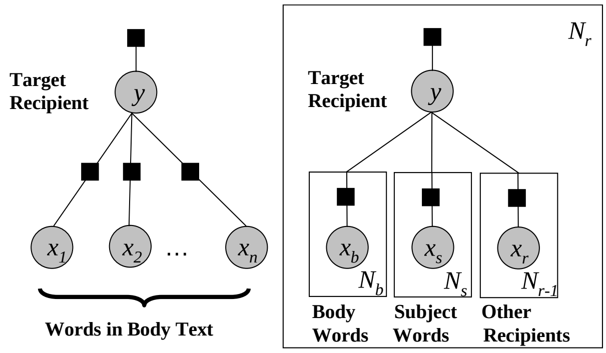

Figure 1: (Left) A factor graph for a naive Bayes model for cc prediction. (Right) A plated factor graph for a naive model employing different alphabets for words in the body, words in the subject line and for recipients.

vocabulary for the body text is 19921 words. The size of the vocabulary for words in the subject line is 2613.

In all our models and experiments we place a Dirichlet prior on the parameters of all the conditional distributions in a model. In practice this amounts to using plus one Laplace smoothing. We evaluate our models using the following procedure. We begin by estimating the parameters of our models on the first week of email and then re-estimate the models each day at 4:00am. We evaluate models by making cc predictions for each email in the sent mailbox over the course of the next day whenever there are multiple recipients. For an email with N_r recipients we generate N_r test cases by removing each of the recipients in turn from the email and making predictions using the remaining recipients as observations. We score a cc prediction as correct if the held out recipient is contained within a list of the top five recipient predicted by sorting the probabilities $p(y|\{x\})$ obtained from the model where $\{x\}$ is the set of all observations in the email and y is a random variable with states corresponding to each possible recipient. Table 1 shows the average daily accuracy for the first and last month and over the timecourse.

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 encoding 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 particular 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 different 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 system [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.

Model	First Month	Last Month	Avg. Daily
Naive Bayes	.301	.326	.364
Factor Graph	.364	.395	**.448**
Thread Info	.357	.403	.448

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

We have used factor graphs with locally normalized factors 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 advantage 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 desirable 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 information 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 organization 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 framework 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 popular approach of introducing hidden topic variables [1] into our model or add more detailed sender and recipient structure 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.

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

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[2] Calo: A cognitive agent that learns and organizes. http://www.ai.sri.com/project/calo, 2006.

[3] F. R. Kschischang, B. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[4] A. McCallum, A. Corrada-Emmanuel, and X. Wang. Topic and role discovery in social networks. *In the Proceedings of 19th International Joint Conference on Artificial Intelligence*, pages 786–791, 2005.

[5] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. *In AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[6] A. McCallum, C. Pal, G. Druck, and X. Wang. Multi-conditional learning: Generative/discriminative training for clustering and classification. In *To appear in AAAI '06: American Association for Artificial Intelligence National Conference on Artificial Intelligence*, 2006.

[7] R. Raina, Y. Shen, A. Y. Ng, and A. McCallum. Classification with hybrid generative/discriminative models. *In NIPS 16*, 2004.

[8] C. Sutton and A. McCallum. An introduction to conditional random fields for relational learning. *To appear in: An Introduction to Statistical Relational Learning. Edited by Lise Getoor and Ben Taskar. MIT Press*, 2006.

[9] B. Wellner, A. McCallum, F. Peng, and M. Hay. An integrated, conditional model of information extraction and coreference with application to citation matching. *In the proceedings of Uncertainty in Artificial Intelligence (UAI), Banff, Canada*, July 2004.

[10] W.R.Gilks, S. Richardson, and D.J.Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice.* Chapman and Hall, 1996.