AN UNSUPERVISED APPROACH TO EMAIL LABEL SUGGESTIONS

David Harwath, Nikhil Johri and Edouard Yin

Abstract

Organization of an email inbox can often become tedious, especially when one receives numerous emails per day. We propose an automatic label suggestion system for email, which uses an unsupervised approach to cluster related emails together based on both latent features, such as the semantics of the email, as well as direct features like the sender and recipients. Our approach utilizes Latent Dirichlet Allocation (LDA), a popular topic modeling technique to determine the topical distributions of each email. These latent distributions are then used as features, along with the more direct features, in the clustering of the emails via k-means clustering. Finally, labels are suggested for each cluster, using its most prominent features, to describe the emails within it.

1. Data collection

Email inbox data was collected in two ways for this task. Two personal email inboxes were downloaded by the authors, consisting of approximately 6,000 and 18,000 emails respectively. Additionally, smaller sized inboxes were taken from the Enron email corpus. We only report tests that have been run on the smaller Enron inboxes.

The data was cleaned using simple regular expressions to remove email headers and unwanted characters, such as forwarded email metadata, hyperlinks and numbers.

2. Topic Modeling

We ran LDA over a number of inboxes of the Enron email dataset. We utilized the LDA implementation of the *Stanford Topic Modeling Toolbox*[1], using the default parameters of the toolbox for 1500 iterations and 30 topics. The sizes of our inboxes varied from 971 emails to 1377. In Table 1, we demonstrate some of the topics we observed. Topic 8 shows a topic relating to financial matters, particularly about banks and lenders, while Topic 24 deals with personal, partly cheerful emails and Topic 28 is a topic about management. Despite running our tests on a very restricted corpus, there are still some clear themes that we can see in output of the topic model.[1]

[1] http://nlp.stanford.edu/software/tmt/tmt-0.3/

Topic 8		Topic 24		Topic 28	
lenders	25.496	i'm	33.81328	ets	142.4442
budget	18.557	make	20.51053	president	135.9027
model	13.803	happy	19.85081	counsel	113.9282
want	12.187	sure	18.50421	general	106.9884
banks	11.644	day	17.07777	vice	101.5829
review	11.457	think	15.11029	mike	97.13978
eca	10.695	hope	14.81352	chairman	95.00147
peter	9.787	thinking	13.98348	report	88.89668
lender	9.074	great	12.8458	director	88.34954
rights	8.775	message	12.82339	currently	77.86481
s&w	8.152	let	12.57449	ceo	71.50873
sponsor	7.900	home	11.62745	organization	68.94505
costs	7.557	good	11.36377	managing	66.77347
contingency	6.920	sue	11.00226	role	65.61392
mary	6.898	lot	10.73718	businesses	63.71853
opic	6.347	love	10.69858	wholesale	62.2071
arrange	6.116	birthday	9.964258	continue	61.7992
support	5.698			business	60.16497
				global	59.73905

Table 1: A sample of the topics discovered by running LDA over an inbox in the Enron email corpus.

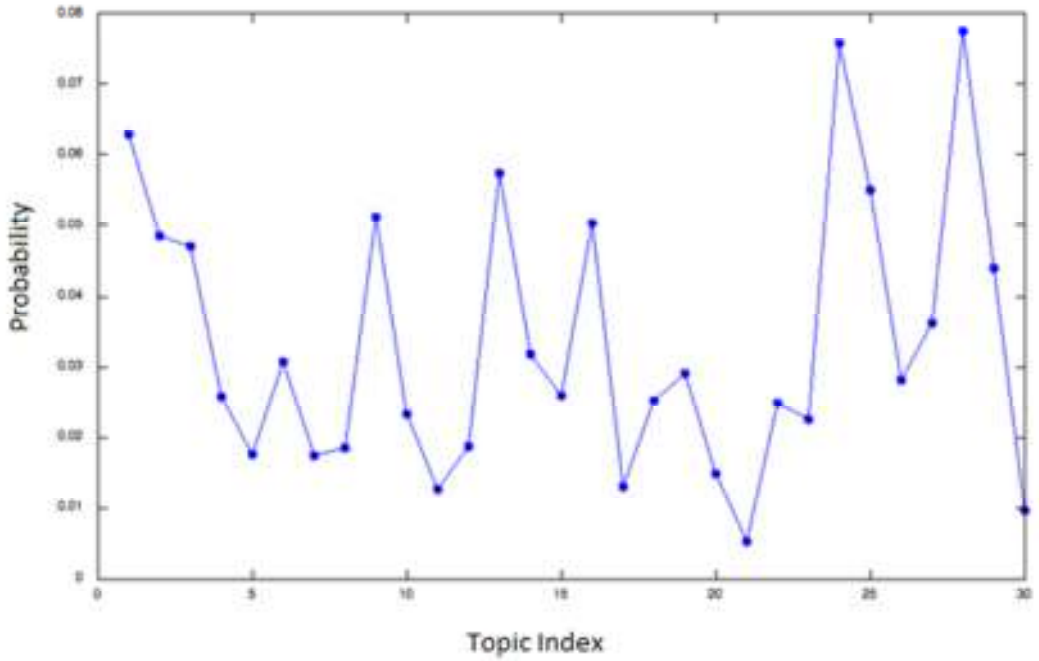

Figure 1: Distribution of topics across all emails

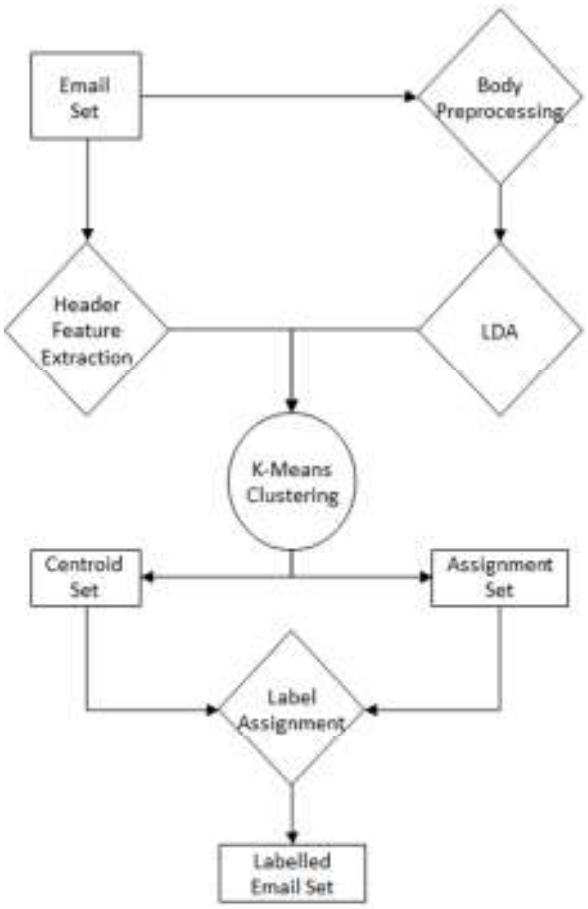

Figure 2: A block diagram view of our system. All testing was done on email inboxes from the publicly available **Enron Email Corpus**.

Figures 1 helps in visualizing the influence of the topic models. We observe that there are about 8 dominant topics out of the 30 found for this particular inbox, and we find these to be among the topics that dominate clusters, invoking label suggestions based on top words of the topics.

3. K-Means Clustering

Once we have the topic distributions, we include them as features for our unsupervised k-means clustering of the emails. We also use, as direct features, the sender of the email, the recipients and the prominent subject words. Additional features that may be interesting to include are whether or not the email was from a mailing list, or whether the email was sent to single or multiple recipients.

We ran our algorithm on several inboxes and our initial results showed that the binary valued features (i.e. sender, recipient, subject words etc.) overpowered the real valued topical weights which had values ranging from 0 to 1. This is to be expected, given that the most dominant topic present in an email will never have a value greater than 1, and will usually have a value between 0.5 and 0.75. However, this is not our desired result, since we would like to cluster based on both semantic topics as well as the direct, explicit features.

Table 2: A sample of the emails clustered by Topic 8 of Table 1 (significant topic words in bold)

Clustered emails with label suggestion 'lenders'
I spoke with **Mary** Mervene this morning and the **lenders** declined to convene this week although we will arrange an **ECA** call for Thursday or Friday. **Mary** has requested that we do the following: 1) Send the **Lenders** and **S&W** the new **budget** info and the revised change order summary. 2) Revise the **model** to …
The **lenders** have declined to meet with us at this time. They have suggested the following instead: 1) Send the **Lenders** and **S&W** the new **budget** info and the revised change order summary. 2) Revise the **model** to reflect the new **costs** and the new timeline. 3) **S&W** will review the **model** and then …
In non-default situations the **Lenders** allow the **banks** to have rights and typically substantial input on waivers modifications etc. As you can imagine the IFC would not want the liability of being responsible to B **banks** if they changed the deal without their consent and the B **banks** would e uncomfortable turning over that right to a third party. They would lose control of the credit process…

To account for this problem, we add a tuning parameter α to our model, whereby the binary features could take the value 0 or α rather than 0 or 1. We set the value of α to 0.5, which we found gave a good balance between clusters dominated by topics and clusters dominated by direct features. However, this value can be altered with respect to the user using the system, as certain users may prefer clustering based on semantic topics while others may prefer explicitly featured topics.

4. Results

We present some sample clustering results in Table 2 and Table 3. The results are based on clusters dominated by topics 8 and 24 from Table 1. As one can clearly see, the emails clustered in Table 2 are all personal emails, usually between the sender and his wife, while those in Table 3 all relate to an issue involving banks and lenders. Our system would provide the top *m* terms from each topic to the user to label these clusters with. An intuitive label for the Table 3 emails, *lenders*, could thus be selected.

Clustered emails from 'personal'-related topic (see Figure 3)
I was reviewing the pictures you **sent**. You are not only an adventuress but also a Goddess. **Happy** Valentines **day**. I turned Pete down on the offer to stay. I guess **I'm** just ready to move on and this feels pretty dead end here. But I'll try and catch up with you from time to time if I can. Beija Rob
I guess you got tied up on some real work. Call me back when you have time. Are you **happy**? The London job sounds **great** and sometimes the best things happen to us unexpectedly. It was nice to hear your voice briefly. I miss seeing you.
Thanks, same to you. I was not here on the **day** so this is late but best wishes. Kathy M Lynn 01/25/2001 04:50 PM To: Rob G Gay/NA/Enron@Enron cc: Subject: **Happy Birthday Happy Birthday**! Please tell everyone that you are 29 so you don't blow my cover since we were born on the same **day**! **Hope** you have a **great birthday**.

Table 3: A sample of the emails clustered by Topic 24 of Table 1 (significant topic words in bold)

Given the subjectivity of email clustering, it was hard to perform a quantitative evaluation. The fact is that different people have individual preferences with respect to the number of desired labels for their inbox, and use different heuristics to assign labels. This means that two people looking at the same inbox may choose vastly different labeling systems. Our evaluation is therefore entirely qualitative, based on the high level cohesiveness and our perception of the quality of the clusters when we read the results.

5. Conclusions

Looking at the output of our system, we came to the following conclusions:

- Email label suggestion is difficult for a machine to do the same way humans do because it would require deep semantic understanding.
- Topic models combined with sender and recipient information do a good job at finding patterns in email collections
- Training data size is very variable, as we want to create a personalized model for each inbox, and inbox size differs from person to person.
- Future work could focus on a better way to weight topical data versus metadata (timestamp, sender address, etc.)
- One possible way to numerically evaluate the quality of the label clusters would be to set up an experiment in which a collection of human subjects run the system on their own personal inboxes, and then give a numerical score to how well they think the system did overall.

6. References

- Blei, D., A. Ng, and M. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research.*
- MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297*
- Ozcaglar, Cagri. (2008). Classification of Email Messages Into Topics Using Latent Dirichlet Allocation: *M.S. Thesis Submitted to Rensselaer Polytechnic Institute, Troy, New York*