Likelihood Prediction of Email Participants using Multi Dimensional Association Networks

Himangshu Ranjan Borah, Sukriti Sharma and Sushma Ravichandran
Computer Science, NC State University

Introduction

The rapidly growing email spectrum has posed some serious limitations on the management of emails. Our research focusses on one such common problem of directing emails to a wrong recipient or overlooking a few important ones.

Our model is novel in the way that it incorporates the different communication relationships between the candidate participants and the sender of the emails for all kinds of probabilistic predictions using association networks.

Communication Graphs

A Generalised Association Network was defined comprising of a TO-Graph, CC-Graph and a BCC-Graph to represent dependencies among different users.

In each of the graphs, $G = (V, E)$, V represents the set of users. The weight of the directed edge (u, v) in edge set E represents the number of emails sent from user u to user v.

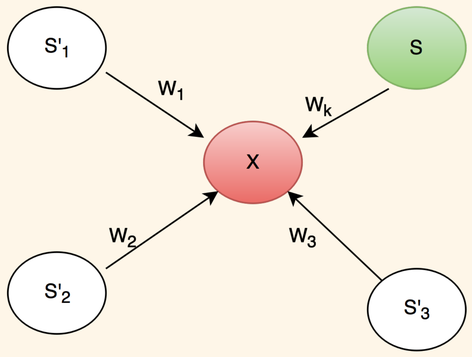

Figure 1: Sample Communication Graph

Probabilistic Modeling

If S is a sender, X a recipient and M is the message content, then we find the probability of X taking a value given the sender S and message content M. Formally, it is $P(X|S, M)$. Using Bayes' Equation and Markov Chain expansion,

$$P(X|S,M) = \frac{P(S|X) \cdot P(M|S,X) \cdot P(X)}{P(S,M)} \qquad P(X|S,M) = \alpha \cdot P(S|X) \cdot P(M|S,X) \cdot P(X)$$

$$P(M|X,S) = \prod_{w \in M} \lambda P(w|X,S) + \gamma P(w|X) + \beta P(w) \qquad P(S|X) = \frac{W_k}{\sum_i W_i}$$

Flow Charts

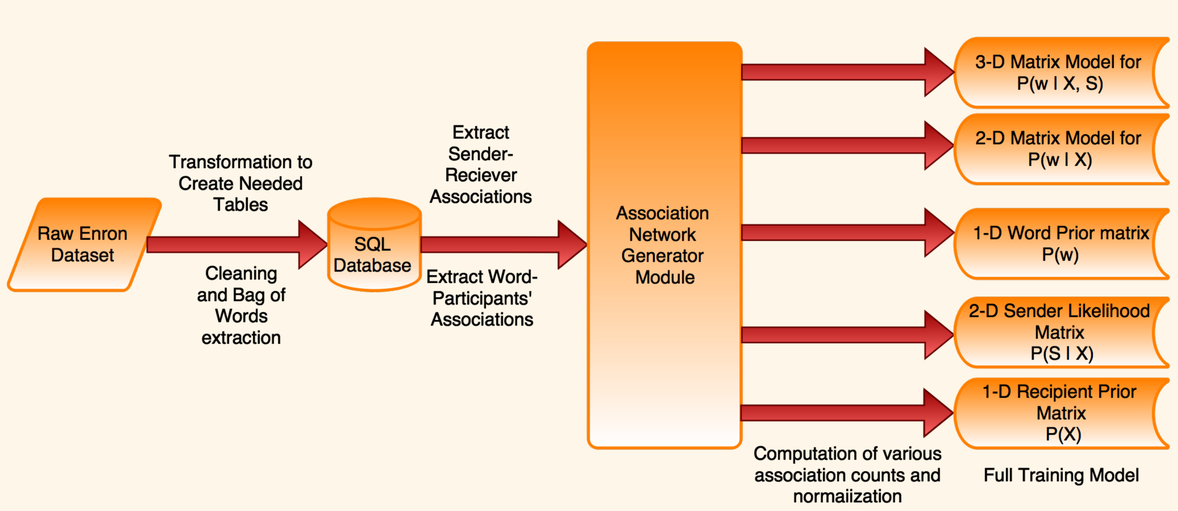

Figure 2: Training Phase

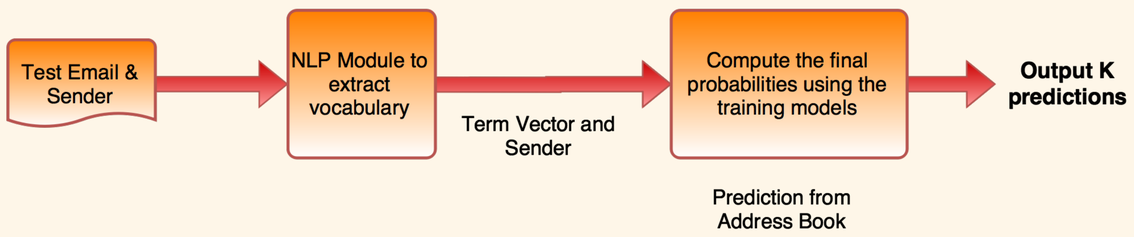

Figure 3: Testing Phase

Experiments and Results

We find MAP, Precision @ 5, Precision @ 10 and R-Precision for the predictions (k = no. of predictions, s = sample size of mails per user),

- Using email content and metadata.
- Using only metadata.

(k, s)	MAP	P@10	R-Prec	P@5
(10, 5)	0.516	0.273	0.273	NA
(10, 10)	0.545	0.392	0.288	NA
(5, 5)	0.523	NA	0.23	0.23
(5, 10)	**0.546**	NA	0.239	0.432

Table 1: Results using content with metadata.

(k, s)	MAP	P@10	R-Prec	P@5
(10, 5)	0.243	0.24	0.072	NA
(10, 10)	0.273	0.254	0.09	NA
(5, 5)	0.236	NA	0.052	0.233
(5, 10)	0.271	NA	0.071	0.258

Table 2: Results using only metadata.

Conclusion

The model performed better than the conventional TFIDF Centroid based approaches which is because it takes sender of the email into consideration unlike the former one. Also it gave much higher scores when incorporated with content as compared to metadata alone. These were the major results that we observed.

The Regularization of the sender likelihood and the dynamic update of the predictions are two of the very promising future works that we have to explore in the coming days.

References

[1] Vitor R Carvalho and William W Cohen. "Ranking users for intelligent message addressing", 2008

[2] Graus, David, David van Dijk, Manos Tsagkias, Wouter Weerkamp, and Maarten de Rijke. "Recipient recommendation in enterprises using communication graphs and email content.", 2014

Dataset

The dataset used in our experiments is a corpus of emails from **ENRON** Corporation available online. The original dataset was preprocessed and modified to form an SQL database and create the necessary tables.