Likelihood Prediction of Email Participants using Directed Association Networks

Himangshu Ranjan Borah hborah@ncsu.edu
Sushma Ravichandran sravich@ncsu.edu
Sukriti Sharma ssharm18@ncsu.edu

Abstract

In this era of technological boom and smart devices, we have started to expect automatization everywhere. Digital communication is one area on which we are becoming increasingly dependent day by day, and Electronic mail or email, is the most important media among them. However, with the increasing number of emails and corresponding participants, we face the problem of participant mis-management in emails. In this paper, we present a generalized architecture of email participants' likelihood prediction which is modeled using association networks between the users as well as rich text mining features extracted from the content of the emails. Our model is novel in the way that it incorporates the different communication relationships between the users for the probabilistic prediction of the candidate participants. We focus on the prediction of all kinds of participant fields like TO, CC and BCC of a new email being composed by a particular sender, which is unlike most of the past works which uses communication networks for this problem. Also, we address a severe problem of sender likelihood prediction when there exists no previously known communication between a pair of users. We combine our predictions with some existing well proven classifiers like TF-IDF Centroid and design an ensemble.

1 Introduction

Emails statistics taken from [10] say that the average business emails sent/received per day is about 100.5 billion by 929 million users. However, the rapidly growing email spectrum has posed serious limitations on the management of emails such as directing an email to a wrong recipient or overlooking a few important recipients. [4] found that 9.27% of the users in Enron dataset had at least one message in which they forgot to add a desired participant. Hence there is a need for accurate architectures to assist users while composing emails.

The critical challenge lies in the fact that the task of email participant prediction is not like the traditional machine learning tasks where we have a matrix format of the data. Due to its complex representation and one-to-many class association structure, we need sophisticated representation of the data. We use association networks for the knowledge representation part and probabilistic modeling for the reasoning part of the system. The results are further boosted by the existing bag-of-words based classifiers by combining evidences from them. We will consider our modeling based on enterprise emails in our paper, however the same concepts can be linearly extrapolated to the general user domain.

2 Related Work

Pal and McCallum in [2] were among the first ones to use the notion of probabilities using multinomial Naive Bayes model in the pursuit of CC field prediction. Later, Carvalho and

Cohen used email content to model TF-IDF Centroid and KNN Classifiers for recipient prediction in [3]. Roth in [5] suggested a paradigm using implicit social graphs which was a pure metadata based approach. Building on that idea, Cohen in [4] mapped the problem of recipient recommendation with expert searching problem. Some of the recent work was done with similar approaches using metadata and modeling them using graphs in [9]. In [10], the authors presented an architecture for this problem where they used groups of recipients. The problem of email participant recommendation shares striking similarities with the problem of email folder classification. We have taken some really useful concepts from this domain to incorporate in our system as found in [7]. Also, for extracting rich features from the email content and the subject of the messages, we used features from text mining domain[11].

3 Generalized Association Networks

We define three different kinds of communication graphs to represent dependencies among users in our approach namely TO-Graph, CC-Graph and BCC-Graph, which we collectively term as Generalized Association Networks somewhat similar to [5][9]. Each of these are directed graphs V = (G, E), where G is the set of all the users involved in the email traffic and the edge set E consists of directed edges (u, v) where "u" sent an email to "v". The arc points from "u" to "v" and the weight of the arc is equal to the no. of emails. Depending on the type of the recipient, "v" can be either of the three categories TO, CC and BCC. We construct three graphs for each type to get our association network. We will use this data structure to compute *Sender Likelihood* in our upcoming modeling section.

4 Probabilistic Modeling

Abductive reasoning has always been a great friend of Machine Learning related problems. In this paradigm, we try to predict the likelihood of some event when we know the information of the occurrence of some of the related events. Mathematically, if S is the random variable for the sender, X is the random variable for the recipient(we will replace X by T for TO, B for BCC and C for CC later) and M be the variable for the message content, then our task is broadly to find the probability of X taking on a particular value given the sender S and message content M. Formally, it is $P(X|S, M)$. We can expand this probability using Bayes Theorem and Markov Chain as follows,

$$P(X|S, M) = \frac{P(S|X) \cdot P(M|S, X) \cdot P(X)}{P(S, M)} \tag{1}$$

Using MAP estimation, we introduced alpha for the denominator as it is the same for all the recipients. We get,

$$P(X|S, M) = \alpha \cdot P(S|X) \cdot P(M|S, X) \cdot P(X) \tag{2}$$

So, the task now boils down to predicting the three different components namely $P(S|X)$, $P(M|S, X)$ and $P(X)$. The first being called the *Sender likelihood*, second one the *Content Likelihood* and the last one is the *Recipient Prior*. Also, we are analyzing the probabilities for three kinds of recipients TO, BCC and CC.

4.1 Likelihood Estimate of Sender

Sender Likelihood ($P(S|X)$) is defined as the probability that S is the sender of a particular mail given X is the recipient of an email. We use generalized association networks for this purpose. A frequency based approach(The number of emails S sent to X in the training period) is taken to count this term which corresponds to the edge weights of the communication graphs. The figure below depicts a basic view for a particular X and S pair. We denote the other users with S'. The weight W_k is the number of emails from S to X and similarly W_i is defined for the S' users.

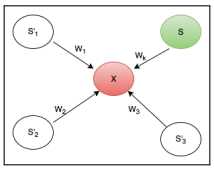

Figure 1: A Sample view of sender likelihood.

Now we define the likelihood as follows,

$$P(S|X) = \frac{\text{Number of emails sent from S to X}}{\text{Total no. of emails received by X from all users}} = \frac{W_k}{\sum_i W_i} \tag{3}$$

4.2 Recipient's Prior

Typically defined as the general probability that a user will receive any message, Recipient's Prior is the total number of emails received by the user normalized by the total email traffic. We use the straightforward naive approach to calculate this. However, we can also use other sophisticated approaches like *PageRank score* as some of the previous work has used like in [9] which tries to find the importance of a recipient X in the association graph.

4.3 Content Likelihood and Bag of Words

Email content likelihood is defined as the probability that a particular email can be observed between a sender and a receiver. We describe each email using a bag of words model. Each document is represented by the terms in the document and the corresponding count of each term. We describe email likelihood as follows[9]: $\lambda + \gamma + \beta = 1$.

$$P(M|X,S) = \prod_{w \in M} \lambda P(w|X,S) + \gamma P(w|X) + \beta P(w) \tag{4}$$

Here, probability of any word in the set, is the frequency of each word in the set by the total number of words in the set.

$P(w|set) = \frac{n(w,set)}{\text{no. of words in the set}}$

5 Database Preparation and Experiments

The dataset used is a corpus of emails from ENRON Corporation available from link. We further modified a data set by Shetty and Adibi in [6] to suit our needs. The original database contained 4 SQL Tables namely *employeelist, message, recipientinfo and referenceinfo*. The major modifications done to the original database are,

- Added new table communications to map the sender and receiver relationships.
- Parsing the whole Email Collections and insertion of unstructured data to DB.
- Splitting the tables into TRAINING and TESTING samples.

All the experiments presented in this paper were implemented using R and python with MySQL community edition open source database server. We sort all the messages by the timestamps and take alternate emails for training and testing set which gives us a stratified sampling of the data according to time. Also we find the highly active users in email sending(more than 100) and receiving(more than 1000) as the candidate set for our classification problem and discard the rest of the users. These decisions are influenced by the work in [9], where they used a similar user pruning criteria to find effective results.

Cleaning and Bag of Words generation

We observed issues in the database version in [6]. The last word of a line and the first word of the next line was getting concatenated, creating new words. We parsed the email contents to eliminate this problem and updated the MySQL database. After doing so, we observed that the bag of words count reduced considerably. Furthermore, special symbols appearing in the content and stop words were removed from the bag of words. To assist in future calculation, each document in the training set is represented as a collection of terms with their corresponding counts. Table 1 shows few term counts per email.

Table 1: Term count matrix

Document Number	Term	Count
0	markets	7
0	services	15
1	enron	11

Prior probabilities of the recipient can be found directly from DB. For the **communications graphs**, we first calculate the count of all the sender-recipient pairs' occurrence and then build the graphs. The top 5 sender-receiver pairs are listed below in Table 2.

Table 2: Top Sender Receiver pairs for "TO" relation

Sender	Recipient	Count
pete.davis@enron.com	pete.davis@enron.com	1245
jeff.dasovich@enron.com	susan.mara@enron.com	713
jeff.dasovich@enron.com	richard.shapiro@enron.com	693
outlook.team@enron.com	no.address@enron.com	662
jeff.dasovich@enron.com	paul.kaufman@enron.com	661

6 Future Experiments

So far, we have calculated two components of the eq. 2 for the training sets and computed the models. We will complete the modeling of the third component to get the final results. We then need to tune the other parameters and report the performances using matrices like *Mean Average Precision(MAP), Mean Reciprocal Rank(MRR)* etc. as done by earlier researchers like in [4][9]. The major remaining experiments,

Expansion of Communication Graphs We will extend the communication networks for all the three kind of recipients like TO, CC and BCC and model all of them.

Sender Likelihood Regularization In the above eq. 3 there's a problem which the previous related work has overlooked. If there's a recipient to whom the sender has never sent a message then the the term $P(S|X)$ becomes zero which forces the whole MAP estimation of the particular recipient to go off from the scenario. This might lead to serious troubles when the other two terms of eq. 2 suggests high probability for that recipient. This problem is similar to the zero probability problem in Naïve Bayes classifier. We try to address this problem in a similar way as *Laplace Transformation* in Naïve Bayes by regularizing eq. 2 as follows,

$$P_{regularized}(S|X) = \frac{W_k + 1}{\sum_i W_i} \tag{5}$$

Dynamically changing predictions After finalizing a few recipients, we can experiment with dynamic update of recipient list. Our previous approach will give us a ranked list of recipients based on sender and email content. Once the user selects a recipient, we want

to explore the social communication of that recipient to determine how it can affect the selection of other recipients. Here X1 is the recipient selected by the user. X2 is a variable for remaining recipients. $P(X2|S, M, X1) = \alpha \cdot P(S|X1) \cdot P(M|S, X1, X2) \cdot p(X2|S, X1) \cdot P(X1)$

Combining Evidences After modeling the system using association networks, we compare the predictions using some existing multi-class classification system and create an ensemble out of it. For this, each email is considered a separate document and the bag of words created by the content of the emails is used to quantify the importance of each term in the data set using a *Term Frequency-Inverse Document Frequency (TFIDF)* score. A technique called *TFIDF centroid* is used to predict the similarity between two TFIDF vector based representations of email messages [3].

Time Dimension We will also try to incorporate time dimension in Sender Likelihood prediction and check if it earns us some extra improvements.

References

[1] Bryan Klimt and Yiming Yang. "The enron corpus: A new dataset for email classification research". In Machine Learning: ECML 2004, pages 217?226. Springer, 2004.

[2] Chris Pal and Andrew McCallum. "Cc prediction with graphical models". In Conference on Email and Anti- Spam, 2006.

[3] Vitor R Carvalho and William Cohen. "Recommending recipients in the enron email corpus". Machine Learning, 2007.

[4] Vitor R Carvalho and William W Cohen. "Ranking users for intelligent message addressing". In Advances in Information Retrieval, pages 321?333. Springer, 2008.

[5] Maayan Roth, Assaf Ben-David, David Deutscher, Guy Flysher, Ilan Horn, Ari Leichtberg, Naty Leiser, Yossi Matias, and Ron Merom. "Suggesting friends using the implicit social graph". In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 233-242. ACM, 2010.

[6] Jitesh Shetty and Jafar Adibi. "The enron email dataset database schema and brief statistical report". Information sciences institute technical report, University of Southern California, 4, 2004.

[7] Tam, Tony, Artur Ferreira, and André Lourenço. "Automatic foldering of email messages: a combination approach." Advances in Information Retrieval. Springer Berlin Heidelberg, 2012. 232-243.

[8] Hu, Qi, Shenghua Bao, Jingmin Xu, Wenli Zhou, Min Li, and Heyuan Huang. "Towards building effective email recipient recommendation service." In Service Operations and Logistics, and Informatics (SOLI), 2012 IEEE International Conference on, pp. 398-403. IEEE, 2012.

[9] Graus, David, David van Dijk, Manos Tsagkias, Wouter Weerkamp, and Maarten de Rijke. "Recipient recommendation in enterprises using communication graphs and email content." In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pp. 1079-1082. ACM, 2014.

[10] Desai, A.; Dash, S.K., "Email recipient prediction using reverse chronologically arranged implicit groups," in Contemporary Computing (IC3), 2014 Seventh International Conference on , vol., no., pp.461-466, 7-9 Aug. 2014

[11] Liu, Tao, Shengping Liu, Zheng Chen, and Wei-Ying Ma. "An evaluation on feature selection for text clustering." In ICML, vol. 3, pp. 488-495. 2003.

Plan of Activities

- **Himangshu :** Parsing Structured Features, Modeling the Association Networks, Calculation of Recipient Prior.(Upcoming : Regularization, Network Expansion)
- **Sushma :** Finding Normalized TF-IDF Score matrices, TF-IDF Centroid implementations(Upcoming : Combining Evidences, Dynamic).
- **Sukriti :** Cleaning the data, Parsing Email Database and Creation of Bag of Words Models.(Upcoming : Content Likelihood, Time Dimension inclusion)