Likelihood Prediction of Email Participants using Multi Dimensional 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, and incorporate multidimensional matrix structures to model the text-participants' associations which is unlike most of the past works which uses communication networks for this problem. The usage of indexing using matrix structures expedited the predictions to a significant extent. Also, we address a severe problem of sender likelihood prediction when there exists no previously known communication between a pair of users. The whole architecture was trained and validated on the Enron Dataset. We compare our predictions with some existing well proven classifiers like TF-IDF Centroid and achieved better results as compared to the baseline publications.

1 Introduction

Emails are one of the most important media for communication when it comes to formal discussions. Some statistics taken from [10] says that currently the average number of business emails are about 100.5 billion, sent/received per day, by 929 million users. These numbers show the exponential growth and usage of emails. However, the rapidly growing email spectrum has posed some serious limitations on the management of the emails. Some of the very common problem of such mismanagement is to direct an email to a wrong recipient or to overlook a few important recipients. In [4] they described that they found 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 dire need of some accurate architectures to assist users while composing their emails. We have couple of existing models in practice today which solves this problem, but none of them seems to be quite accurate and hence it is a very active area of research in machine learning.

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 incorporating content based information from the emails and 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

An email has normally two kinds of data associated with it, namely structured data and unstructured data. The former refers to the metadata like participants' emails ids, date/time etc. and the latter corresponds to the raw text data that appears in the subject and body of the emails (often referred to as the "content" of the email). Almost all of the previous approaches for email classification related tasks deal with either structured or unstructured data alone, or a mix of both. The problem of email participants prediction is sometimes called as a recipient recommendation problem as well. Among the first ones to try this problem, the work of Pal and McCallum in [2] is a notable one. They used the notion of probabilities using multinomial Naive Bayes model in the pursuit of CC field prediction. After that, Carvalho and Cohen published some of the great work in this area. Their work was among the first ones to take email content into consideration to model TF-IDF Centroid and KNN Classifiers for recipient prediction in [3]. Another class of approaches considered social communication networks for the participant recommendation problem. For the first time, Roth, in [5] suggested this paradigm using implicit social graphs which was a pure metadata based approach. Building on that idea, Cohen published one more legendary work in [4] where he mapped the problem of recipient recommendation with expert searching problem in organizations. Some of the recent work was done with similar approaches using metadata and modeling them using graphs in [9]. In [10], they presented a unique architecture for this problem where they used groups of recipients to find the most probable sets. The problem of email participant recommendation shares striking similarities with the problem of email folder classification. We have taken some really useful and nice 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 or BCC. We construct three graphs for each type to get our association networks. We will use this data structure to compute *Sender Likelihood* in our upcoming modeling section.

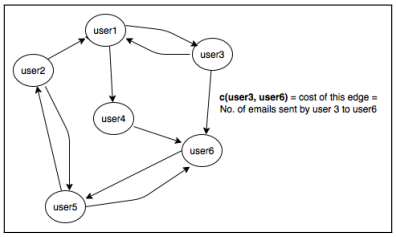

Figure 1: A Sample Association Network.

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.

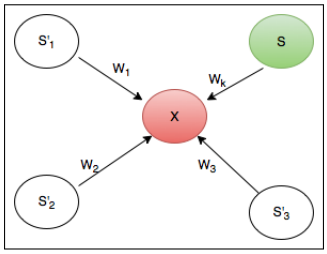

Figure 2: A Sample view of sender likelihood.

Now we define the likelihood as follows,

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

Sender Likelihood Regularization

In the above eq. 3 there's a problem which the previous related work has overlooked. Assuming a closed world like an organization, 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,

$$P_{regularized}(S|X) = \frac{W_k + 1}{\sum_i W_i} \tag{4}$$

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]: $\lambda + \gamma + \beta = 1$. For out experiments, we use the values found in [9] which they claim to give best performance in these scenario.

$$P(M|X,S) = \prod_{w \in M} \lambda P(w|X,S) + \gamma P(w|X) + \beta P(w) \tag{5}$$

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.

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

4.3.1 Representing Content Likelihood as Multi Dimensional Matrices

Due to the huge size of the vocabulary, it's very important to represent the content-user relationships efficiently. To completely eliminate linear search and counting, we model multi dimensional matrix structures to find the 3 terms of the eq. 5 during the training phase. These structures are computed to find the co-occurrence frequency of the terms and the users which is extracted from the email database prepared from the original raw dataset(To be described in the next section). For eg., to compute $P(w|X,S)$, we make a 3-dimensional matrix to compute the counts of messages where the term w, sender S and receiver X appear together. We found during our experiments, that using these structures speeded up our predictions to a significant extent as compared to naive lazy learning methods as uses direct indexing to retrieve the values(from the order of second to milliseconds). However, computing these structures in training phase takes up relatively more amount of time as compared to direct counting in runtime, but we believe that reducing prediction time is much more important that reducing training time. Also, using these structures, updating of the models become pretty straight forward. Also, we used similar structures for modeling the association graphs mentioned in section 3.

The whole architecture of the system is shown in Fig. 3 and Fig. 4 as Training and Testing Phases.

5 Plan

The main goal of our project is to experiment the effects of incorporating probabilistic modeling for the recipient recommendation problem. We designed the preliminary model for the testing phase of recipient prediction using only Sender Likelihood and Recipient Prior. We proceeded by incorporating the content based information extracted from the

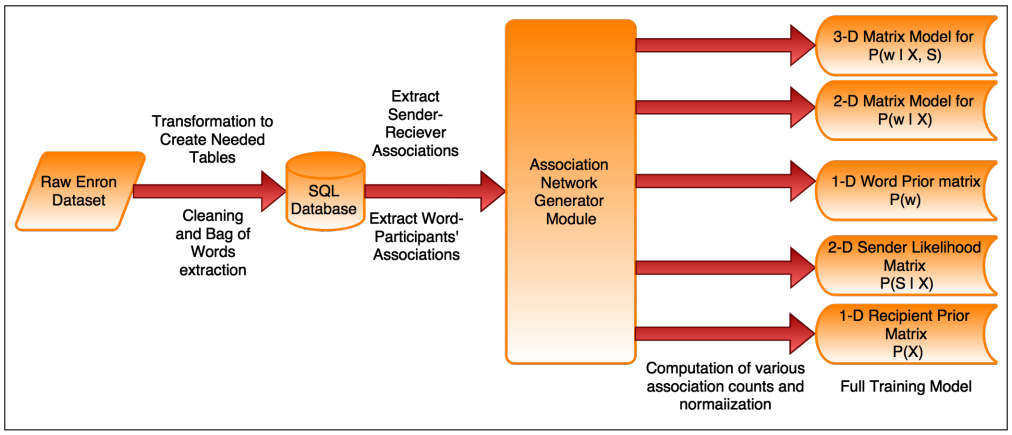

Figure 3: The Training phase.

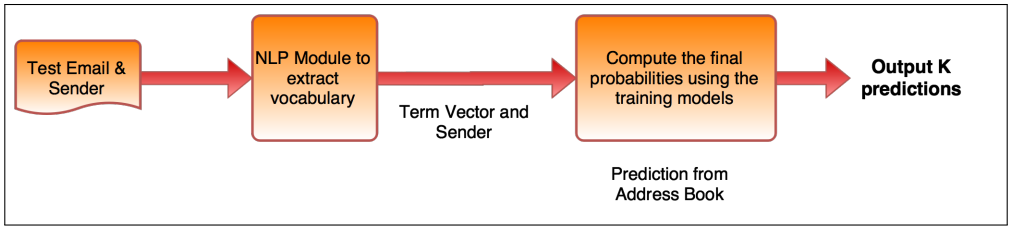

Figure 4: The Testing phase.

emails. We plan to contrast the results obtained from both sets of predictions and analyze the importance of textual information in the emails. Apart from this, we intend to train and test our data set using the concept of TF-IDF Centroid[3] which solely uses the textual content of emails to predict recipients. Finally, we compare these outcomes with those of our model.

6 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 a new table *communications* to map the sender and receiver relationships.
- Parsing the whole Email Collections and insertion of unstructured data to DB.
- Inclusion of a new table *document_term* to represent the content based data for 500 most relevant words.
- 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 (sending more than 100 or receiving more than 1000) as the candidate set for our classification problem and discard the rest of the users. We finalize a total of 41 senders and 3319 recipients. These decisions are influenced by the work in [9], where they used a similar user pruning criteria to find effective results. We also introduce the concept of a *Sender Address Book(AB)* and a *Global Address Book(GAB)*. The AB of a sender S consists of all the users to whom S has sent

at least one mail. An the GAB consists of all the possible recipients of the environment. However, the concept of GAB is more useful and relevant when we have a closed world assumption like an organization.

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. Lemmatization, a technique of transforming different forms of a word into a single one, was also applied to the term in the email. To assist in future calculation, each document in the training set is represented as a collection of terms with their corresponding counts.

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. For the content likelihood modeling, we compute the matrix structures as described before.

We present the recipient recommendation experiments and results using the models described above. The traditional classifier evaluation methods like accuracy, precision etc. are not relevant in this kind of recommendation systems due to it's inherent properties. To evaluate our model, we compute more elegant matrices like R-Precision, Precision at 5 , Precision at ten, and mean average precision on the test set[4]. We first evaluate our model that uses only structured features (metadata without content). We then evaluate our model that uses both structured and unstructured features (metadata and content). We design test sets by varying the number of predictions to be made per email (k) and varying the sample size of emails per user (s) . We used random sampling to select the emails. The test sets and results have been shown in the below tables. The results are averaged over 41 users and 3319 recipients in total.

(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.

During our project, we observed many interesting results about the dynamics of the recommendation system that we modeled.

One of the major observations was that the results were considerably better for the model that used both content and structured data than the model that used only structured data. This proves that the email content plays a major role in recommending recipients. Also the results for each of these evaluation measures increases with increase in sample size from 5 to 10 for the same k value. However, there is one interesting behavior that we observed with different k values. Intuitively, increasing the no. of predictions should increase the probability of good hits. But it was not true. For example, in case of s = 5, k = 10 gives an MAP of 0.516 where k = 5 gives an MAP of 0.523. Investigating , we found that having

included more predictions, the rate of increase of irrelevant recipients is much higher than the relevant ones. As MAP penalizes the wrong predictions equally, it brings down the overall performance. This is a good indictor not to show the user with more recommendations as it might lead to an annoying user experience in the long run. We believe that such indicators will help us in analyzing the behavior of recommendation systems and help us designing more efficient ones.

TFIDF Centroid TFIDF Centroid is a technique proposed by [3] which takes advantage of the frequency of terms in any users' emails. Each email is treated as a document and the TFIDF vectors for each document is generated from its textual contents. For each of the 3319 recipients r_i in the data set, a TFIDF centroid vector is obtained by summing up the normalized TFIDF vectors of all messages that were sent to r_i.

We compared our results with the above TFIDF Model and the values that we obtained are pretty satisfactory as compared to the baseline results of TFIDF enumerated in the previous works like [3], with a maximum MAP of 0.546 for the k = 5 and s = 10 scenario.

7 Conclusion and Future Work

Any insights gained during a project can be usefully applied on future projects. Some of the significant observations made by us, from the results as well as during the course of coding have been stated below.

We learnt that both association analysis of the user and the textual content of the email are important factors in Recipient Recommendation Systems. While implementing our model, we learnt the importance of indexing in MySQL database for faster retrieval of data. However, we found that retrieving from the MySQL database was still very time consuming, so we had to implement the efficient matrix structures for the modeling of the content likelihood as described above in detail.

Also we learnt about the concepts of the rich features of Natural Language Processing domain and the elegant evaluation metrics of rank based retrieval systems. The experience with the terms in the emails also taught us the importance of dimensionality reduction, since using the entire set of bag of words was increasing the time complexity of our algorithm. The overall learning experience was great during the execution of the project.

Due to insufficient time for exhaustive experimentation, some of the very interesting add-ons that we had planned for had to be left out. Among many other small improvements in the modeling section and implementation part, 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. The first one was already discussed in the mathematical modeling section. We did try few predictions for testing the effect of incorporating Regularization, but we were not able to conclude much from those as they need more extensive and rigorous experiments. Technically, if we speak in terms of sender likelihood and content likelihood, we have to find out appropriate test scenarios in which the sender likelihood was initially zero but the content likelihood was very high for a particular recipient who is not there in the sender's AB. Having found such a test mail, we can quantitatively prove the claims about it.

Another very interesting future work that we feel is about Dynamically changing predictions in the recommendation system. 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. $P(X2|S, M, X1) = \alpha \cdot P(S|X1) \cdot P(M|S, X1, X2) \cdot p(X2|S, X1) \cdot P(X1)$

Finally, we have to extend all the results for both CC and BCC fields which is just a horizontal expansion of the modeling that we have described. We look forward to implement and test the above scenarios in coming future which will be a big leap forward in this recipient recommendation problem domain.

Acknowledgement

An endeavour such as this one, in order to reach fruition, requires the contribution of a variety of sources, whose selflessness and painstaking efforts have enabled this undertaking to reach its best possible level. *Dr. Min Chi*, Computer Science Department, NC State University, has been of great support as the course instructor as well as the project guide for the $CSC522$. Her inputs and helping hand were essential to create the best possible version of the project. We are humbled to have worked with Teaching Assistants $Yuan$ and $Krishna$ who as our guide have been an incredible support to us and a wealth of knowledge with regard to the project.

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.