MODELING INTERACTIONS FROM EMAIL COMMUNICATION

Dong Zhang[1], *Daniel Gatica-Perez*[1], *Deb Roy*[2] *and Samy Bengio*[1]

[1]IDIAP Research Institute and Ecole Polytechnique Federale de Lausanne (EPFL) Switzerland
[2]Massachusetts Institute of Technology, Cambridge, MA 02142, USA

ABSTRACT

Email plays an important role as a medium for the spread of information, ideas, and influence among its users. We present a framework to learn topic-based interactions between pairs of email users, i.e., the extent to which the email topic dynamics of one user are likely to be affected by the others. The proposed framework is built on the influence model and the probabilistic latent semantic analysis (PLSA) language model. This paper makes two contributions. First, we model interactions between email users using the semantic content of email body, instead of email header. Second, our framework models not only email topic dynamics of individual email users, but also the interactions within a group of individuals. Experiments on the Enron email corpus show some interesting results that are potentially useful to discover the hierarchy of the Enron organization.

1. INTRODUCTION

Email has become one of the most important media for human communication. It is indispensable in organizations for both local and remote information sharing and collaboration. Several properties distinguish email from other media: (i) semi-structure: structured header ("To", "From", "Date") and unstructured body (the text of the email); (ii) sequential nature: every email has a timestamp (date); (iii) plentiful data in electronic form; (iv) possibly multimedia email attachments: People not only exchange text (in both email subject and body), but also e-documents (papers, pictures, weblinks, presentations ...). In this sense, email is truly multimedia.

There has been increasing interest in email research, mainly in social network analysis (SNA) [11]. Previous work on emails has been limited by two factors: (1) unavailability of a public corpus from a real organization; (2) privacy issues: only "To" and "From" fields of emails have been used, ignoring the email content. The Enron email corpus (publicly available at `http://www-2.cs.cmu.edu/∼enron/`) is appealing not only because it is a large scale email collection from a real organization covering a period of 3.5 years, but also because it uniquely documented the rise and fall of the energy giant Enron. It provides a promising resource for research on human interactions, and for discovery of the hidden patterns of collaboration and relationships in communities.

There has been quite recent work on the Enron corpus. Most work has focused on natural language processing (NLP) perspectives, such as spam detection and email topic classification [4, 8]. The exploration of both NLP and SNA has started with the author-recipient-topic model (ART) [9], a static Bayesian network, investigating the use of email content to discover roles of the people in the social network. To our knowledge, however, little work has been conducted to study the influence between email users, while the problem of determining how much influence one person has on others has been studied using other media, such as video and audio, in a number of settings, e.g., multi-party conversations [3], and wearable computing [6].

In this paper, we propose a framework that qualitatively investigates the interaction and influence among email users. The proposed framework is built on the influence model [3] and probabilistic latent semantic analysis (PLSA) [7]. This paper makes two contributions: (i) Instead of using email traffic ("From" and "To" fields), we model interactions between emails users using the semantic content of emails. (ii) The proposed framework uses a dynamic Bayesian network (DBN) to model not only email topic dynamics of individual email users, but also the interactions within a group of individuals. Although the Enron corpus used in our experiments only contains text, our framework could be easily applied to multimedia e-mail corpus . For instance, in case there were image attached in emails, both textual and visual PLSA features [10] could be extracted and serve as inputs to the influence model.

The paper is organized as follows. Section 2 gives an overview of the proposed framework. Section 3 presents email topic modeling using PLSA, and Section 4 describes the influence model. An agglomerative clustering is described in Section 5. Section 6 reports the results on the Enron dataset, and an email visualization and retrieval system. In Section 7, we discuss the limitations of our framework, and present future directions.

2. FRAMEWORK OVERVIEW

Our framework (Fig. 1) includes several parts. First, an email parser automatically extracts the standard email items, i.e., *sender, recipient, subject, date,* and *the body* from the email text file. Second, we perform standard text preprocessing on

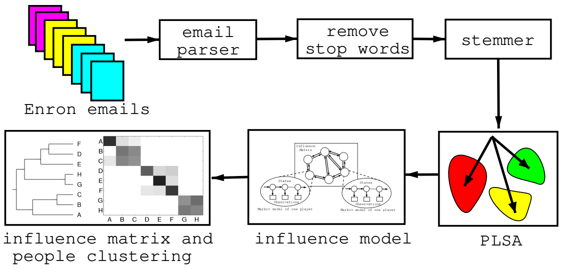


Fig. 1. The proposed framework to learn influence among people from emails.

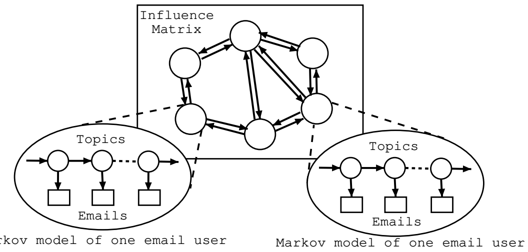


Fig. 2. Influence model. The model has two levels. The first level models email topic dynamics of individual users, and the second level models interactions within a group of individuals.

the email body, including removing stop words, and stemming word using Porter's suffix-stripping algorithm. Thirdly, we apply PLSA language model [7] to project each email from the high-dimensional bag-of-words space into a low-dimensional topic-based space (Section 3). The output of PLSA serves as input to the influence model, which learns how much influence each email user has on the others (Section 4). The learned model is an influence matrix in which each entry α_{ij} represents the influence of person i on person j. The degree of interaction between two persons is defined as the average of the pairwise influence: $\beta_{ij} = \frac{1}{2}(\alpha_{ij} + \alpha_{ji})$. A clustering algorithm can be applied to the interaction matrix to cluster people into groups for the discovery of the community structure of the organization (Section 5). More details will be described in the following sections.

3. MODELING TOPICS WITH PLSA

Probabilistic latent semantic analysis (PLSA), also called aspect model, is a language model that transforms documents in the high-dimensional bag-of-words space to a low-dimensional topic-based space. Each dimension in this new space represents a topic, and each document is represented as a mixture of the topics. In our case, a document corresponds to one email. We summarize the PLSA model in the following. For a detailed discussion, see [7].

In PLSA, the conditional probability between documents d and words w is modeled through a latent variable z, which can be thought of as a topic. A PLSA model is parameterized by $P(w|z)$ and $P(z|d)$. It is assumed that the distribution of words given a topic, $P(w|z)$, is conditionally independent of the document. Thus the joint probability of a document d and a word w is represented as

$$P(w, d) = P(d) \sum_{z} P(w|z)P(z|d). \tag{1}$$

The PLSA parameters, $P(w|z)$ and $P(z|d)$, are estimated using the EM algorithm to fit a training corpus D with a vocabulary of W, by maximizing the log-likelihood function

$$L = \sum_{d \in D} \sum_{w \in W} f(d, w) \log P(d, w), \tag{2}$$

where $f(d, w)$ is the frequency of word w in document d.

Starting from random initial parameter values, the EM procedure iterates between:

- **E-step:** where the probability that a word w_j in a particular document d_i is explained by the topic z_k is estimated as:

$$P(z_k|w_j, d_i) = \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)}. \tag{3}$$

- **M-step:** where the parameters $P(w_j|z_k)$ and $P(z_k|d_i)$ are re-estimated to maximize L in Equation (2):

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} f(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{j=1}^{M} \sum_{i=1}^{N} f(d_i, w_j)P(z_k|d_i, w_j)}, \tag{4}$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{M} f(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{k=1}^{K} \sum_{j=1}^{M} f(d_i, w_j)P(z_k|d_i, w_j)}, \tag{5}$$

where N is the number of documents in the corpus D. M is the number of words in the vocabulary W, and K is the number of PLSA topics. The EM iterations are stopped once the relative difference in the global log likelihood is less than 2%.

Given the learned PLSA model, we can transform each email into a K-dimension vector ($K = 50$ in our experiments), in which each dimension gives the probability of the email belonging to each of the topics.

4. THE INFLUENCE MODEL

We describe the structure and learning of the influence model in this section. The full motivations and justifications were originally described in [2].

4.1. Model Structure

The influence model (Fig. 2) is a dynamic Bayesian network (DBN) that models interacting Markov chains. The entire network has a two-level structure: the individual user level and the interaction level. For the individual level, we model email topic dynamics of each email user using a first-order Markov model with one observation variable and one state variable. In our case, the observations are emails, and the states represent the topics conveyed by emails. To model interactions, the state at time t of the user i (S_t^i) depends on all the previous states of all users (including itself i), resulting in the full conditional state transition probability: $P(S_t^i|S_{t-1}^1 S_{t-1}^2 \cdots S_{t-1}^N)$, where N is the total number of persons.

The influence model [2, 3] employs the strategy that reduces the full conditional probability as a convex combination of pairwise conditional probabilities,

$$P(S_t^i|S_{t-1}^1 S_{t-1}^2 \cdots S_{t-1}^N) = \sum_{j=1}^{N} \alpha_{ji} P(S_t^i|S_{t-1}^j), \quad (6)$$

where α_{ji} ($\sum_{j=1}^{N} \alpha_{ji} = 1$) represents how much the state transition the i^{th} Markov chain is influenced by the j^{th} Markov chains. In other words, α_{ji} represents the influence of person j on person i, corresponding to the weight of the link from i to j of the influence matrix (Fig. 2). Note that $\alpha_{ij} \neq \alpha_{ji}$, i.e, the influence of person i on person j is not equal to the influence of person j on person i. The interaction between person i and j can be defined as $\beta_{ij} = \frac{1}{2}(\alpha_{ij} + \alpha_{ji})$, which is used as the similarity between a pair of persons to cluster people into groups (Section 5).

4.2. Learning the Influence Matrix

The maximum likelihood (ML) criterion can be applied to estimate the model parameters. The joint log probability of the influence model is

$$\begin{aligned} \log P(S,O) = & \underbrace{\sum_{i=1}^{N} \log P(S_1^i)}_{initial\ probability} + \underbrace{\sum_{t=1}^{T}\sum_{i=1}^{N} \log P(o_t^i|S_t^i)}_{emission\ probability} \\ & + \sum_{t=2}^{T}\sum_{i=1}^{N} \log \sum_{j=1}^{N} \underbrace{\alpha_{ji} P(S_t^i|S_{t-1}^j)}_{j\ influence\ on\ i}, \end{aligned} \quad (7)$$

where O and S denote observations and states respectively. T is the length of the sequence, and o_t^i denotes the observation of person i at time t. Similar to the aspect HMMs [5], we embed PLSA as the emission probability in Equation (7), which means that we have K (the number of topics in PLSA) different states for the variable S_t^i. In [3], the gradient descent was used to calculate the α_{ji} values by maximizing Equation (7). We keep only the terms relevant to maximization over α_{ji} in Equation (7),

$$\alpha_{ji}^* = \arg\max_{\alpha_{ji}} \{\sum_{t=2}^{T}\sum_{i=1}^{N} \log \sum_{j=1}^{N} \alpha_{ji} P(S_t^i|S_{t-1}^j)\}. \quad (8)$$

Taking the derivative with respect to α_{ij}, we get,

$$\frac{\partial \log P(S,O)}{\partial \alpha_{ji}} = \sum_{t=2}^{T}\sum_{i=1}^{N} \frac{P(S_t^i|S_{t-1}^j)}{\sum_{j=1}^{N} \alpha_{ji} P(S_t^i|S_{t-1}^j)}. \quad (9)$$

More details are given in [3].

5. CLUSTERING PEOPLE

As discussed in Section 4, the learning result of the influence model is the interaction matrix, in which each entry of row i column j (β_{ij}) tells us the degree of interaction between person i and j. Motivated by the assumption that interactions among people in the same group are usually strong, and interactions among people in different groups are normally weak, we apply a standard agglomerative clustering method on the interaction matrix, described as follows. We start with each person forming its own cluster, and iteratively merge clusters which have the largest interaction value until all people have been gathered into a single big cluster. The similarity of two clusters is calculated as the average of the pairwise interaction of the persons from each cluster. That is, $Sim(C_i, C_j) = \frac{1}{N_i N_j} \sum_{k \in C_i, l \in C_j} \beta_{kl}$, where N_i, N_j is the number of persons in cluster C_i and C_j, respectively. β_{kl} is the interaction between person k (in cluster C_i) and person l (in cluster C_j).

6. EXPERIMENTS

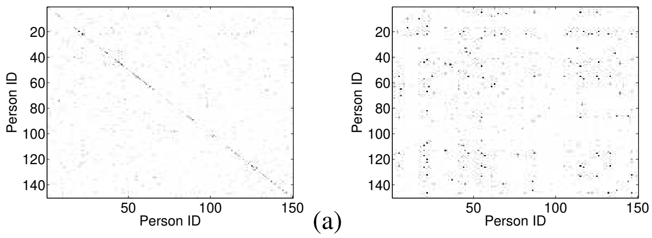
In this section, we first briefly describe the Enron corpus and the data preprocessing, then present our results. Finally, we briefly describe our email visualization and retrieval system with the feature of user clustering.

6.1. Enron Corpus and Preprocessing

The Enron email dataset was made public by the US Federal Energy Regulatory Commission (FERC) during its investigation into Enron affairs. The cleaned version contains 517,431 messages sent by 150 personnel of the corporation between 1998 and 2002 [8]. In our experiments, we only used the emails that were received by at least one of the 150 users, amounting to 21,612 emails. The 21,612 emails were ordered according to their date with a time step of one day from Oct. 13, 1998 to May 21, 2002. The PLSA topic for the day without emails was set to zero, and multiple emails in the same day by the same person were merged. After applying language preprocessing including downcase, removal of the stop words, and word stemming, we obtained a vocabulary of 23,776 unique terms.

6.2. Results

Fig. 3 (a) shows the learned interaction matrix. The value of each entry of row i column j (β_{ij}) is the interaction between person i and person j. As a comparison, we calculated



(a)
(b)

Fig. 3. (a) The interaction matrix. (b) The email traffic matrix.

Table 1. Statistics of interaction and the email traffic matrix.

Matrix	min	max	mean	std.
Interaction	0	0.9931	0.0067	0.0356
Traffic	0	7102	4.72	86.69

another matrix based on the email traffic between users. In specific, the weight of the link between user i and user j is the number of emails between i to j, denoted by M_{ij}. The M_{ij} matrix, which we call the email traffic matrix, is shown in Fig. 3 (b).

We can see that both matrices are symmetrical and sparse, but the interaction matrix has a clear diagonal (β_{ii}), which indicates the email topics of most users are influenced by their own Markov dynamics. Table 1 shows some basic statistics of the two matrices, including the *min value, max value, mean value*, and the *standard deviation*. Table 2 lists some examples of the pairwise interaction (β_{ij}) and the number of emails between two persons (M_{ij}). The table items are listed based on β_{ij} in descending order. We can see that a large M_{ij} may not correspond to a large β_{ij}. For example, the number of emails of pair D: "Jeff Dasovich" and "Mary Hain" is 248, which is larger than that of pair B: "Teb Lokey" and "Steffes Corman" (37). But the interaction estimated by our approach of pair D (0.012) is much smaller than that of pair B (0.28). This might be explained by their job titles. The job titles of pair B were both related to regulatory affairs, while pair D had quite different roles in the organization: one is the government relation executive and one is a lawyer. Similar reasons might explain the other items in the Table. We can see that

Table 2. Examples of the pairwise interaction (β_{ij}) and the number of emails between two persons (M_{ij}). The job titles were found using google search.

Pair	person i		person j		β_{ij}	M_{ij}
	name	job title	name	job title		
A	Jeff Dasovich	Government Relation Executive	James Steffes	Vice President Government Aff.	0.49	1182
B	Teb Lokey	Manager Regulatory Aff.	Shelley Corman	Vice President Regulatory Aff.	0.28	37
C	Jeff Dasovich	Government Relation Executive	Steven J. Kean	Chief Staff Government Aff.	0.16	172
D	Jeff Dasovich	Government Relation Executive	Mary Hain	In-house lawyer	0.012	248
E	Stanley C. Horton	CEO of Gas Pipeline	Rod Hayslett	CFO and Treasurer	0.001	65

β_{ij} is in better accordance with role similarities than M_{ij}.

7. LIMITATION AND FUTURE WORK

The lack of a comprehensive evaluation and comparison with other methods is a typical issue in SNA [11], and also the main limitation of our work. In contexts where researchers know what the right answer should be, evaluation is done by comparing automatic results with the manual ground-truth. In other contexts, evaluation is more subjective because there is no one right answer. Our initial evaluation thus far has used google search for job titles of email users. For a formal and comprehensive evaluation in the future, we have plans for consultations with Enron experts who could identify interesting and useful results.

Another limitation of our approach is the first-order Markov assumption used in the influence model to model topic dynamics of individual email users. Some emails will invalidate this assumption. To handle this, we could use a higher-order Markov model by adding longer temporal dependencies. This will be investigated in future work.

Acknowledgments

This work was supported by the Swiss National Center of Competence in Research on Interactive Multimodal Information Management (IM2), and the EC projects AMI (Augmented Multi-Party Interaction, pub. AMI-172) through the Swiss OFES.

8. REFERENCES

[1] J. Allan, J. Callan, K. Collins-Thompson, B. Croft, F. Feng, D. Fisher, J. Lafferty, L. Larkey, T. N. Truong, P. Ogilvie, L. Si, T. Strohman, H. Turtle, and C. Zhai. The Lemur toolkit for language modeling and information retrieval. `http://www.cs.cmu.edu/~lemur`.

[2] C. Asavathiratham. The influence model: A tractable representation for the dynamics of networked markov chains. *Ph.D. dissertation, Dept. of EECS, MIT, Cambridge*, 2000.

[3] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland. Learning human interactions with the influence model. *MIT Media Laboratory Technical Note No. 539*, 2001.

[4] R. Bekkerman, A. McCallum, and G. Huang. Automatic categorization of email into folders: Benchmark experiments on enron and SRI corpora. *UMass CIIR Technical Report IR–418*, 2004.

[5] D. Blei and P. Moreno. Topic segmentation with an aspect Hidden Markov Model. *Proc. of the 24th International ACM SIGIR conference on Research and development in information retrieval*, 2001.

[6] T. Choudhury and S. Basu. Modeling conversational dynamics as a mixed memory markov process. *Proc. of Intl. Conference on Neural Information and Processing Systems (NIPS)*, 2004.

[7] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. In *Machine Learning, 42:177–196*, 2001.

[8] B. Klimt and Y. Yang. The Enron corpus: A new dataset for email classification research. *European Conference on Machine Learning*, 2004.

[9] A. McCallum, A. Corrada-Emmanuel, and X. Wang. Topic and Role Discovery in Social Networks. *In IJCAI*, 2005.

[10] F. Monay and D. Gatica-Perez. On Image Auto-Annotation with Latent Space Models. in Proc. ACM Multimedia , Berkeley, Nov. 2003.

[11] J. Tyler, D. Wilkinson, and B. Huberman. Email as spectroscopy: automated discovery of community structure within organisations. *in Communities and Technologies*, 2003.