International Journal of Security and Its Applications
Vol. 9, No. 4 (2015), pp. 229-238
http://dx.doi.org/10.14257/ijsia.2015.9.4.21

Unsupervised Extraction of Signatures and Roles from Large-Scale Mail Archives

Yuan Xiaoqin

Beijing International Studies University
yuan@bisu.edu.cn

Abstract

In this paper, we focus on the problem of signature and role extraction from large-scale mail archives. Due to the huge scale and great diversity of large-scale mail archives, the extraction methods should not only be able to extract signatures and roles accurately without any training data, but also be general enough to work well with large-scale mail archives with different characteristics. To address this problem, we first propose an unsupervised language model based method to identify sig-natures from large numbers of emails, and then present an unsupervised two-stage method to effectively extract roles from the identified signatures. Experimental results on two real-world datasets show that our methods are general and effective for both the signature and role extrac-tion from large-scale mail archives.

Keywords: *Mail Archives, Signature Extraction, Role Extraction, Language Model*

1. Introduction

Email has become one of the most important means of communication in collaborative workgroups, especially in large organizations. The growing importance of email has inspired many attempts to develop intelligent tools for classifying, organizing and presenting email messages. Signatures and roles are two important components in emails, and are very useful for characterizing the senders of the emails. Signatures often consist of personal information of senders, such as their job titles and contact information. Roles are often contained in signatures to identify the senders' job titles, or their relations to some important entities, such as projects and products. Signatures and roles are both carefully designed by the senders to communicate their identity, so are important in email processing. For large-scale mail archives in large organizations, they are more important because they can precisely characterize the expertise, the contact information, and the relationships to projects or products of a large number of organization members, which are very important for applications like expert finding [1] and e-discovery [9]. In this paper, we focus on the effective and efficient extraction of signatures and roles from large-scale mail archives, to support these applications.

Previous study [3] has focused on the extraction of signature lines from emails, where some manually selected features are created for each line, and then some supervised learning methods are used to train classifiers to identify whether each line is a signature line. This method has achieved effective results on a small email corpus, but cannot be generalized and applied to large-scale mail archives. The supervised learning methods require a high-quality training set that reflects the characteristics of all the emails to be processed. For large-scale mail archives, due to the diversity in style, format, and content of the emails, and the prohibitively high cost of tagging the equally diverse training corpuses, this approach has difficulty in scaling to large-scale mail archives. Furthermore, even if we can obtain an equally diverse training corpus, the trained classifiers are only effective on the similar email corpus. To address these problems, we need an unsupervised and general method which does not require training data, and can extract signatures from different mail archives effectively.

Roles in emails have also attracted the interests of many researchers [8, 10,4]. [8] finds that different roles correspond to different distributions of speech act types [5] in emails, which can be used to predict the roles of the senders given some known senders' roles. [10] utilizes the topic distributions of senders to predict their roles from several known senders' roles, with the assumption that the senders with the same role have similar topic distributions. [4] proposes a method combining traffic-based and text-based email patterns to predict the leadership role for each sender. All of these studies focus on the prediction of the roles of senders on the basis of some known senders' roles. However, for large-scale mail archives which involve a large number of senders, we would need to manually identify enough senders' roles, which is impractical. Moreover, the roles in these studies, such as "leader" and "professor", are too general to be useful in enterprise applications. In our research, we aim at automatically extracting the detailed roles from large-scale mail archives. The extracted roles can not only be used to predict the roles of other senders, but also to support other applications such as expert finding and e-discovery.

The challenge of effective extraction of signatures and roles lies in two aspects. First, large-scale mail archives have large numbers of emails with different senders, which belong to different departments, and often design their signatures and specify their roles using different styles and terms. Therefore, to perform effective extraction, our methods should be general enough to accommodate the great diversity of the signatures and roles in large-scale mail archives. Second, due to the huge scale of mail archives in large organizations, the cost of creating training datasets which can reflect the characteristics of the signatures and roles in large numbers of emails, might be prohibitively high. As a result, our methods should be able to support effective extraction with no training data.

To overcome the above challenges, we propose two unsupervised methods. For the extraction of signatures, we first make use of the sender's name to identify the beginning of the signature in each email, and then propose a language model based method to identify the ending of the signature. This unsupervised method does not require training, and can accommodate the diversity in large-scale mail archives. After that, we propose a two-stage method to extract roles from signatures. First, we build a position-based language model on the identified signatures to ensure that the words related to roles get larger probabilities, and then use the model to identify the candidate roles. Second, we build a role language model on the candidate roles, and use it to distill the candidate roles by identifying the missed roles and replacing the wrong candidates with the right roles. This two-stage method is completely unsupervised and can extract roles effectively. Finally, we demonstrate our technique on two real large-scale mail archives, and the experimental results show that our technique is general and effective.

The rest of this paper is further structured as follows. The next section discusses related work. Section 3 describes the problem formulation. Section 4 proposes the unsupervised signature extraction method. Section 5 presents the two-stage role extraction method. Section 6 illustrates the experiments and analyzes the experimental results. And section 7 summarizes this paper.

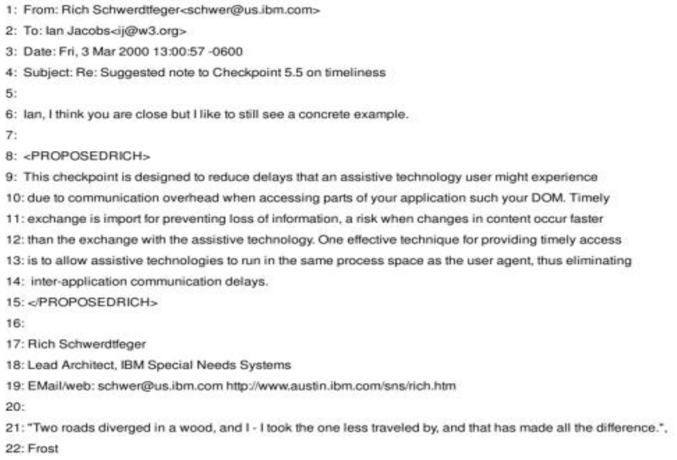
2. Related Work

As the work of this paper focuses on the extraction of signatures and roles from large-scale mail archives, the related work lies in two aspects: signature extraction and role prediction. We will introduce the previous studies in these two areas.

Previous studies on signature extraction [2, 3] aim at automatically identifying the signature lines in emails. In [2], some simple heuristics are utilized to identify signature lines, and no evaluation result is reported. [3] proposes a supervised method to transfer the signature extraction into a classification problem. Specifically, [3] proposes more than

25 types of features to represent each line in emails as a feature vector, and leverages a manually labeled dataset to learn some statistical models as the classifiers. It also compares the performance of different statistical models and finds that, when representing each line as a vector of the features of itself and its previous and next lines, CPerceptron [6] is the best choice. However, the method in [3] cannot be generalized and applied to large-scale mail archives, because it requires high-quality training sets that characterize the features of all the emails in the archives. Due to the diversity of genres of the large numbers of emails, the cost of tagging such training sets is prohibitively high. Moreover, the trained classifier on one training set is only effective on a similar email corpus. Different from [3], our proposed unsupervised method can not only extract signatures from the large numbers of emails without training data, but also process different mail archives with different characteristics effectively.

The prediction of roles is to predict the roles of senders given some known senders' roles. [8] finds that different roles correspond to different distributions of speech act types [5] in emails, and proposes to predict the roles of the senders using the roles of other senders with similar speech act type distributions. [10] proposes an Author-Recipient-Topic (ART) model to discover the topic distribution of each sender, and finds that the senders with the same roles have similar topic distributions. It then utilizes the topic distributions to predict the roles of senders given several known senders' roles. [4] investigates how team leadership roles can be inferred from a collection of email messages exchanged among team members, and proposes a method which combines traffic-based and text-based email patterns to predict the leadership position for each sender. Neither of these studies takes into account the problem of extracting roles from emails, but assumes that the roles of some senders are known in advance. This implies that we need to manually identify enough sender roles, which is impractical for large-scale mail archives with a large number of senders. Moreover, the roles in these studies, such as "leader" and "professor", are too general to be useful in enterprise applications. Different from these methods, our proposed method aims at extracting detailed roles from large-scale mail archives.



```
1:  From: Rich Schwerdtfeger<schwer@us.ibm.com>
2:  To: Ian Jacobs<ij@w3.org>
3:  Date: Fri, 3 Mar 2000 13:00:57 -0600
4:  Subject: Re: Suggested note to Checkpoint 5.5 on timeliness
5:
6:  Ian, I think you are close but I like to still see a concrete example.
7:
8:  <PROPOSEDRICH>
9:  This checkpoint is designed to reduce delays that an assistive technology user might experience
10: due to communication overhead when accessing parts of your application such your DOM. Timely
11: exchange is import for preventing loss of information, a risk when changes in content occur faster
12: than the exchange with the assistive technology. One effective technique for providing timely access
13: is to allow assistive technologies to run in the same process space as the user agent, thus eliminating
14:  inter-application communication delays.
15: </PROPOSEDRICH>
16:
17: Rich Schwerdtfeger
18: Lead Architect, IBM Special Needs Systems
19: EMail/web: schwer@us.ibm.com http://www.austin.ibm.com/sns/rich.htm
20:
21: "Two roads diverged in a wood, and I - I took the one less traveled by, and that has made all the difference.",
22: Frost
```

Figure 1. An Example Email from the W3C Email Corpus

3. Problem Formulation

Our research focuses on the problem of extracting signatures and roles effectively from large-scale mail archives. To describe the problem clearly, we first define a concept "signature" as follows:

Definition 1 A signature is a set of sequential lines in an email, which identifies the sender's personal information, such as job title, organization, contact information, and the relationship to other important entities. The signature often begins with the sender's name or nickname. If we denote an email $\mathrm{E} = \{\mathrm{l_i}\}_{\mathrm{i=1}}^{\mathrm{N}}$, where $\mathrm{l_i}$ is the ith line in E, its signature is $\mathrm{S} = \left\{\mathrm{l_j}\right\}_{\mathrm{j=m}}^{\mathrm{n}},1 < m < m \ll N$

For example, Fig. 1 illustrates the content of an email $\mathrm{E} = \{l_i\}_{i=1}^{22}$ selected from W3C email corpus [7], with the signature $\mathrm{S} = \left\{l_j\right\}_{j=18}^{19}$.

Definition 2 A role is a line in an email signature, and is used to identify the sender's job title, or relationship to other important entities, such as projects or products. That is, for an email E with signature $\mathrm{S} = \left\{l_j\right\}_{j=m}^{n}$ and role $r,\ \exists i,j \in [m,n], r = l_j$.

For example, the 18th line in Fig. 1, *"Lead Architect, IBM Special Needs Systems",* is the signature which identifies the sender's job title.

It's possible that the sender's role consists of multiple lines of the signature in an email, but it's unusual and we don't take into account it. It's also possible that the sender's role does not appear in the signature, but is stated elsewhere. This situation is related to entity relation extraction [11] and beyond the scope of this paper.

Based on these two definitions, we can formulate our problem as follows: Given a mail archive $A = \{E_i\}_{i=1}^{M}$of a large organization, for each email $E_i$with the signature and role, automatically extract its signatureS_i , and identify the sender's role r_i from the elements in S_i .

4. Unsupervised Signature Extraction

As discussed in section 3, the signature of an email often begins with the sender's name or nickname. We use the sender's name and nickname to identify the beginning of the signature. Specifically, we first generate the sender's nickname using a pre-complied list of person names with corresponding nicknames, and then segment the email into lines to search for the last line which contains the sender's name or nickname. If this line only contains the sender's name or nickname, we identify its next line as the beginning of the signature; otherwise, we remove the sender's name or nickname from this line, and regard it as the beginning of the signature.

After that, the next task is to discover the ending of the signature. For each email, we first define a *"raw"* signature simply by regarding the lines from the identified beginning line to the end of the email as the signature. This raw signature usually contains some noise. For instance, the 21th and 22th lines in Fig. 1 are the noise which needs to be removed. We propose a language model based method to remove the noise. Specifically, we build a language model on the raw signature collection to capture the distribution of the signature- related words. Formally, for a raw signature collection D_{rs}, we build an unigram language modelθ_{rs} , in which the probability $\mathrm{P}(\omega_i|\theta_{rs})$ for word ω_i is estimated by MLE (maximum likelihood estimation) [12] as follows:

$$\mathrm{P}(\omega_i|\theta_{rs}) = \frac{c_{D_{rs}}(\omega_i)}{\sum_j c_{D_{rd}}(\omega_j)} \qquad (1)$$

where $C_{D_{rs}}(\omega_j)$is the occurrence count of ω_i in D_{rs}.

Although the raw signatures often contain some noisy information, the noise is not the major component. Moreover, the signatures are carefully designed by the senders, and the senders in the same organization often tend to use similar words to describe their signatures. As a result, the above language model assigns larger probabilities to the signature-related words than hose of the noisy words,and can be used to differentiate the signatures from noise.

After constructing the signature language model, we leverage it to compute the generative probability for each line in each raw signature. Specifically, given the ith line $l_i = \omega_{l,n}$ in a raw signature, we compute the generative probability of l_i as:

$$P(l_i|\theta_{rs}) = P\big(\omega_{1,n}\big|\theta_{rs}\big) = P(\omega_1|\theta_{rs)}P(\omega_2|\omega_1,\theta_{rs}) \ldots P(\omega_n|\omega_{1,n-1},\theta_{rs}) \quad (2)$$

Assuming the word occurrences are independent of one another, we can compute it as:

$$\mathrm{P}(l_i|\theta_s) = \prod_{j=1}^{n} P(\omega_j|\theta_{rs}) \quad (3)$$

Each raw signature can be divided into two parts, the signature part and the noise part. We believe that the generative probabilities of the lines in the noise part are significantly smaller than those in the signature part. Moreover, in each part, the differences of the generative probabilities between adjacent lines in terms of the orders of magnitude might change slightly, but the difference between the last line of the signature part and the first line of the noise part is more significant. In other words, if we draw a curve to characterize the changes of the differences of the generative probabilities between adjacent lines in terms of the orders of magnitude in the raw signature, the most prominent spike with a positive value always corresponds to the ending of the signature.

For example, Fig. 2 illustrates the change of the generative probabilities of the lines in a raw signature, and we can see that the generative probabilities from the 1st line to the 8th line are relatively larger, but the generative probabilities from the 9th line to the 13th line are much smaller and near zero. Fig. 3 illustrates the corresponding decrease in terms of the orders of magnitude at each line, and we can easily find that the curve has the most prominent spike at 8th point, which corresponds to the boundary of the signature part and the noise part.

Another point need to be addressed is that not all the raw signatures have noisy information. Therefore, only using the above method might cause some errors. To address this problem, we add a constraint to the boundary line that its difference of the generative probabilities from the next line in terms of the orders of magnitude must be large enough to ensure that it’s really the boundary line.

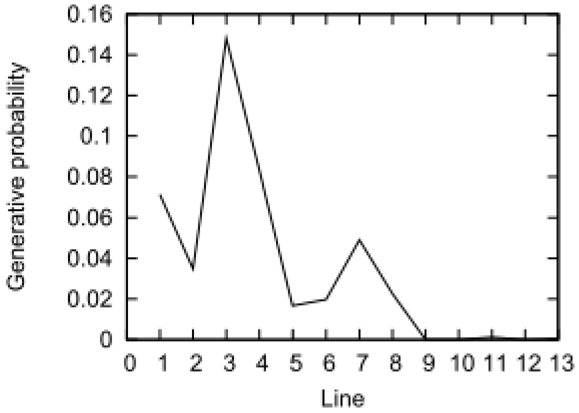


Figure 2. The Generative Probability of Each Line

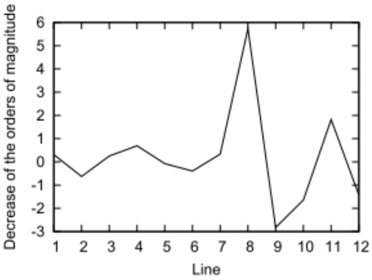


Figure 3. The Decrease of the Orders of Magnitude at Each Line

Consequently, for each raw signature$S_r = \{l_j\}_{j=m}^{n'}$, we can generate its sequence of generative probabilities $< P(l_m|\theta_{rs}), \dots P(l_{n'}|\theta_{rs})$. From this sequence, we want to discover the line $\mathrm{l_n}$, which can satisfy the following three conditions:

$$m < n < n'$$

$$l_n = {}^{argmax}_{\quad l_j}(log_{10}(P(l_j|\theta_{rs})) - log_{10}(P(l_{j+1}|\theta_{rs})))$$

$$log_{10}\big(P(l_n|\theta_{rs})\big) - log_{10}(P(l_{n+1}|\vartheta_{rs}) > \alpha \times \frac{\sum_{j=m}^{n'-1} | \Big(P\big(l_j\big|\theta_{rs}\big)\Big) - log_{10}(P(l_{j+1}|\vartheta_{rs})}{n' - m}$$

where α is a predefined parameter, and we set it to 3.2 experimentally.

If we can find the line satisfying the above conditions, we identify it as the ending of the signature. Otherwise, we identify the last line of the raw signature as the ending of the signature. This way, we can correctly identify the signature, and move on to extract the role from it.

5. Unsupervised Role Extraction

The goal of role extraction is to identify the line describing the role of the sender from the lines in each signature. For this problem, a natural choice is to manually select a set of words identifying roles (such as “manager” and “director”), and use them to identify the lines containing them as roles. However, different organizations usually use different words to describe roles. And for a large organization, the number of role words is very large and the words change overtime.

Therefore, this method is impractical for the extraction of roles from large-scale mail archives. To address this problem, we propose a two-stage unsupervised method to automatically identify the role line from each signature. At the first stage, we construct a position-based language model on the signature collection to ensure that the role-related words are assigned larger probabilities than others, and compute the generative probability of each line in a signature. We then select the line with the largest probability as the candidate role. At the second stage, we build a role language model on the candidate role collection, and use it to distill the candidate roles by selecting the line with the largest probability as the desired role. We will describe the details of these two stages in the following.

5.1 Candidate Role Identification

When people design their signatures in emails, they tend to first describe their roles, and then describe other information such as organizations and addresses. In other words, the roles tend to be put in front of other lines in signatures. However, the roles are not always the first lines of the signatures. Actually, many roles are in the second, third or other lines of the signatures. Based on this observation, we build a position-based language model on the signature collection to ensure that the words related to roles get larger probabilities than others. Specifically, for a word w i in the signature collection$C_s = \{S_q\}_{q=1}^{n}$, we estimate its probability as

$$\mathrm{P}(\omega_i|\sigma) = \frac{PC_{C_s}(\omega_i)}{\sum_j PC_{C_s}(\omega_j)}$$

where $\sigma$is the estimated position-based language model, $PC_{C_s}(\omega_i)$ is to measure the occurrences of ω_i in different lines in the signatures, and can be calculated as

$$C_{C_s}(\omega_i) = \sum_{q=1}^{n}\sum_{k=1}^{|S_q|} C(\omega_i, l_k) \times \frac{1}{k^2}$$

where l_k is the kth line in the signature inS_q , and $C(\omega_i, l_k)$ is the occurrence count of ω_i in l_k.

From this equation, we can see that the words frequently appearing in the front of the signatures will get larger probabilities, and we believe that they are related to roles with larger possibilities. As a result, we can compute the generative probability for each line using the language model, filter the lines with probabilities smaller than a predefined threshold, and select the line with the largest generative probability from the remaining lines as the candidate role.

5.2 Role Distillation

The candidate roles identified in the first stage are not always right. To distill them, we need to identify the missed roles, and find the wrong roles and replace them with the right ones. Although the candidate roles are not always right, we believe that most of them are right. As a result, we can use the statistical information of the candidate roles to perform the distillation. Specifically, we first build a role language model η based on the candidate role collection C_R , in which the probability of the word ω_i is estimated as follows:

$$\mathrm{P}(\omega_i|\eta) = \frac{C_{C_R}(\omega_i)}{\sum_j C_{C_R}(\omega_j)}$$

Where $C_{C_R}(\omega_i)$ is the occurrence count of the word ω_i in C_R .

Obviously, if most of the candidate roles are right, the role-related words will be assigned larger probabilities in this model. With the assumption that the word occurrences are independent of one another, we can utilize this model to compute the generative probability for each line l_j in a signature as follows:

$$\mathrm{P}(l_j|\eta) = \prod_{i=1}^{|l_j|} P(\omega_i|\,\eta)$$

We then filter the lines with probabilities smaller than a predefined threshold, and select the one with the largest probability as the desired role.

6. Experiments

The proposed unsupervised signature and role extraction methods have been evaluated on two real datasets. We conduct a set of experiments to verify the effectiveness and scalability of them.

6.1 Experimental Design

Datasets We select two real-world datasets to conduct experiments. One is the dataset D_{Com} , which consists of 122, 282 mails segmented from mail archive webpages in the Intranet of a large IT company. The emails in this dataset record the communications between sales and engineers, and contain large numbers of signatures and roles. The other is the public W3C email dataset D_{W3C}[7],which consists of 174, 299 emails and also contains large numbers of signatures and roles. These two datasets cover two typical kinds of large-scale mail archives, and are good for evaluating the scalability of our technique.

Baseline Selection As discussed in section 2, the work in [3] utilizes a supervised method to extract signature lines. We select the best performance method *CPerceptron* in [3], and implement it as the baseline to compare the performance with our unsupervised signature extraction method.

Baseline Selection As discussed in section 2, the work in [3] utilizes a supervised method to extract signature lines. We select the best performance method *CPerceptron* in [3], and implement it as the baseline to compare the performance with our unsupervised signature extraction method.

Evaluation Metrics We use three standard metrics, Precision, Recall and F 1 ,to evaluate the performance of extraction. Specifically, Precision is the number of correct extractions divided by the number of extracted items (signatures or roles), Recall is the number of correct extraction divided by the number of existing items, and F 1 is the harmonic mean of Precision and Recall.

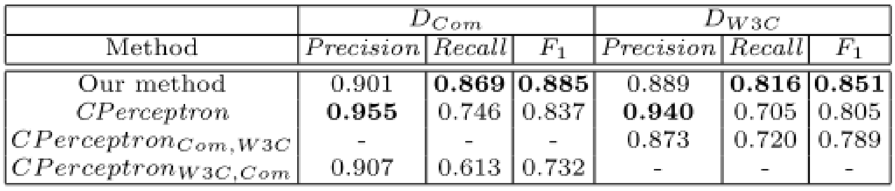
6.2 Experimental Results

Signature Extraction For the dataset D_{Com} , we randomly select 628 emails with signatures, and manually label each line in each email to indicate whether it's a signature line. In all the 20, 820 lines in the 628 emails, 4, 568 lines are labeled as signature lines. For the dataset D_{W3C} , we randomly select 646 emails with signatures, and manually label 2, 514 lines in all the 42, 386 lines as signature lines.

For the proposed unsupervised signature extraction method, we first utilize it to extract signatures from all the emails in each dataset. The results are actually the signature blocks in the emails. To compare the performance of this method with the baseline, we further segment the identified signature blocks into signature lines, and compute the evaluation metrics on the labeled emails in the two datasets.

For the baseline *CPerceptron*, we use 5-fold cross-validation on the two labeled datasets to evaluate its performance. Moreover, in order to discover the scalability of the baseline, we first train two *CPerceptron-based* classifiers using the labeled emails of the two datasets, respectively; we then use one classifier to extract signatures from the labeled emails in the other dataset, and evaluate its performance.

Table 1. The Performance of the Signature Extraction Methods



	D_{Com}			D_{W3C}		
Method	*Precision*	*Recall*	F_1	*Precision*	*Recall*	F_1
Our method	0.901	**0.869**	**0.885**	0.889	**0.816**	**0.851**
CPerceptron	**0.955**	0.746	0.837	**0.940**	0.705	0.805
$CPerceptron_{Com,W3C}$	-	-	-	0.873	0.720	0.789
$CPerceptron_{W3C,Com}$	0.907	0.613	0.732	-	-	-

Table 1 illustrates the performance of the methods on the two datasets. In this table, $CPerceptron_{com,W3C}$ is the method which trains classifier on the labeled emails in D_{Com} , but tests on the dataset D_{W3C} ; And $CPerceptron_{W3c,com}$ is the method training on D_{W3c}but testing on D_{Com} . From this table, we can see that our unsupervised method outperforms the baseline *CPerceptron* in the metrics of *Recall* and F_1 , and the values of its *Precision* are slightly smaller than those of the baseline. This is because our method is able to utilize the statistical information of the whole set of emails in each dataset, which are more abundant than the information in the much smaller labeled email set. We can also see that both$CPerceptron_{com,W3C}$and $CPerceptron_{W3c,com}$ cannot achieve the same performance as $CPerceptron$, and the performance of $CPerceptron_{W3c,com}$ is more worse. This indicates the classifier trained on one dataset cannot be scaled to another dataset.

Role Extraction For the 628 labeled emails in D_{Com} , we manually annotate the role line in the signature of each email, and find that 583 (92.83%) signatures have role lines. We then calculate the ratio of each line being the role line in all the 583 signatures. We also perform the same annotation and calculation on the 646 labeled emails in the dataset D_{W3C} , and find that 203 (31.42%) signatures have role lines.

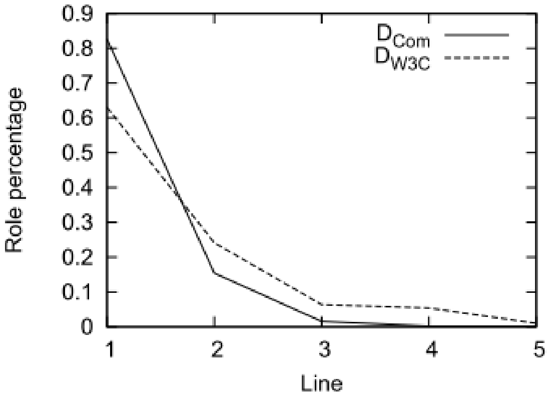
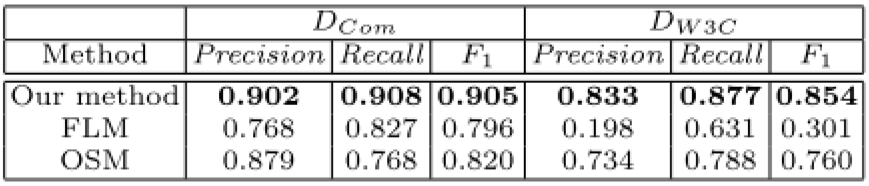


Figure 4. The Curve of Each Line and its Ratio of being the Role Line

Fig. 4 illustrates the ratio of each line being role line. We can see that the first line has the largest probability being role line in both datasets. However, the first line in D_{Com} is dominant with ratio 0.827, but the first line in D_{W3C} (0.631) is not the dominant and no dominant line exists because all the ratios are smaller than 0.7. This indicates that different mail archives have different distributions of role lines, and simply regarding some line as the role line is ineffective.



	D_{Com}			D_{W3C}		
Method	*Precision*	*Recall*	F_1	*Precision*	*Recall*	F_1
Our method	**0.902**	**0.908**	**0.905**	**0.833**	**0.877**	**0.854**
FLM	0.768	0.827	0.796	0.198	0.631	0.301
OSM	0.879	0.768	0.820	0.734	0.788	0.760

Table 2. The Performance of the Role Extraction Methods

To evaluate the performance of the proposed unsupervised role extraction method, we perform role extraction on all the emails in each dataset using our method and the baselines (FLM and OSM). We then compute the evaluation metrics on the annotated datasets. Table 2 illustrates the performance of the role extraction methods on the two datasets. We can see that our method outperforms the two baselines on both datasets, and the values of its F 1 on the two datasets are 0.905 and 0.854 respectively, which indicates that our method has enough scalability to perform extraction on different large-scale mail archives. We can also see that the second stage in our proposed role extraction method, role distillation, can not only discover the wrong candidate roles and replace them with right ones, but also identify the missed roles and increase recall.

7. Conclusion

This paper studies the problem of unsupervised extraction of signatures and roles from large-scale mail archives, which aims at effectively extracting signatures and roles of senders from a large number of emails without any training data. Due to the huge scale and great diversity of large-scale mail archives, it is difficult to support effective extraction from different large-scale mail archives with different characteristics without any training data. We propose two unsupervised methods to address this problem. First, a language model based method is presented to identify signatures from large-scale mail archives. Then, a two-stage method is proposed to effectively extract roles from the large number of identified signatures. Both methods need no training data, and are able to perform extraction on large-scale mail archives with significant diversity. Our experimental results on two real datasets validate the effectiveness and scalability of our proposed methods in the extraction of signatures and roles from large-scale mail archives.

References

[1] K. Balog, L. Azzopardi and M. de Rijke, “Formal models for expert finding inenterprise corpora”, Proceedings of SIGIR, ACM, **(2006)**; New York, NY,USA.

[2] K. Balog and M. de Rijke, “Finding experts and their eetails in e-mail corpora”, Proceedings of WWW, ACM, **(2006)**; New York, NY, USA.

[3] V. Carvalho and W. Cohen, “Learning to extract signature and reply lines fromemail”, Proceedings of the Conference on Email and Anti-Spam, **(2004)**.

[4] V. Carvalho, W. Wu and W. Cohen, “Discovering leadership roles in email work-groups”, Proceedings of the Conference on Email and Anti-Spam, **(2007)**.

[5] W. Cohen, V. Carvalho and T. Mitchell, “Learning to classify email into ‘speechacts’”, Proceedings of the Conference on Empirical Methods on Natural LanguageProcessing, **(2004)**.

[6] M. Collins, “Discriminative training methods for hidden markov models: Theoryand experiments with perceptron algorithms”, Proceedings of the Conference onEmpirical methods in natural language processing, Association for Computational Linguistics Morristown, **(2002)**; NJ, USA.

[7] N. Craswell, A. de Vries and I. Soboroff, “Overview of the trec enterprisetrack”, TREC Conference Notebook, **(2005)**.

[8] A. Leuski, “Email is a stage: discovering people roles from email archives”, Proceedings of SIGIR, ACM, **(2004)**; New York, NY, USA.

[9] V. Luoma, “Computer forensics and electronic discovery: The new managementchallenge”, Computers & Security, vol. 25, no. 2, **(2006)**, pp. 91–96.

[10] A. McCallum, X. Wang and A. C. Emmanuel, “Topic and role discoveryin social networks with experiments on Enron and academic email”, Journal of Artificial Intelligence Research, vol. 30, **(2007)**, pp. 249–272.

[11] S. Sarawagi, “Information extraction”, Foundations and Trends R ° in Databases, vol. 1, no. 3, **(2007)**, pp. 261–377.

[12] C. Zhai, “Risk minimization and language modeling in text retrieval”, PhD thesis, University of Massachusetts, Amherst, **(2002)**.

Authors

Yuan Xiaoqin