Automatically Locating Salutation and Signature Blocks in Emails

Meijuan Yin, Junyong Luo, Ding Cao, Xiaonan Liu and Mingtao Li
Zhengzhou Information Science and Technology Institute
Zhengzhou 450002, China

***Abstract*—This paper focuses on the problem of automatically locating salutation and signature blocks in the body of plain-text emails. Texts of salutation and signature block in an email usually contain identity information about the email's sender or recipients. The analysis of locating and extracting salutation and signature blocks from emails has many potential applications, such as entity attributes extracting, person entity based email social network analysis, anonymization of email corpora, improving automatic content-based email classifiers and email threading. Our approach is based on the statistical method and the rules restricted method, which can greatly improve the locating efficiency and at the same time promise a relatively high accuracy of the extracted blocks. We use the statistical method to roughly estimate the number of lines in salutation and signature blocks, and introduce some restriction rules to refine the lines located by the statistical method. Results on the public subset of the Enron corpus prove the high performance of our approach with the average F1 value above 94%.**

Keywords-email body analysis; loacating; salutation blocks; signature blocks;

I. Introduction

The popularity of the Internet makes email become an important communication medium, which has inspired many attempts to develop intelligent tools for parsing, organizing, and classifying email messages[1][2][3]. And recently some researchers take a deeper study on emails for resolving personal name references and modeling entity's identity in emails[5][6]. For those tools and researches, it is necessary and important to be able to analyze the body of an email message by splitting it into components.

This paper focuses on the problem of automatically locating salutation and signature blocks in the body of plain-text emails. The salutation and signature block in an email usually include some identity information about the email's sender and recipient. The analysis of locating and extracting salutation and signature blocks from emails has many potential applications, such as entity attributes extracting, person entity based email social network analysis, anonymization of email corpora, improving automatic content-based email classifiers, and email threading. Our own interest stems from the need to preprocess email messages in order to extract personal name aliases from email bodies.

As the language used in emails tends to be diverse, informal, and (particularly for signature blocks) often even creative, it is difficult to analyze email bodies in the component-level[9]. We exploit statistical and rules restricted methods to automatically directly locating and extracting salutation and signature blocks from emails, which can greatly improve the locating efficiency and at the same time promise a relatively high accuracy of the extracted blocks. We use the statistical method to roughly estimate the number of lines of salutation and signature blocks in an email, and then introduce some restriction rules to refine the lines located by the statistical method. Results on the public subset of the Enron corpus prove the high performance of our approach with the average F1 value above 90%.

The remainder of this paper is organized as follows. Section 2 reviews earlier approaches related to analyzing the body of an email message. The proposed method to automatically salutation and signature blocks from email bodies is introduced in detail in section 3. In section 4, our method is evaluated on the public subset of the Enron collection. Results and limitations of the approaches are discussed in section 5.

II. Related Work

Our approach to locate salutation and signature blocks in emails relates to the analysis on the body of mail messages. Most of researches on email body analysis are about email body splitting, name reference resolution and entity's identity modeling in emails[4][7][8]. Only few of them have studied the difficult problem of locating the salutation or signature blocks from email bodies.

T. Elsayed et al.[6] addressed the problem of resolving personal name references in the full email including the message body. The method of T. Elsayed et al. used only blank lines to partition the content and signature in the body of an email, which is very easily to be reproduced but effect only for emails with normal bodies. When the body of an email is not consistent with the common format, the method only using blank lines does not work well.

Chen et al.[10] described a method to parse signature block fields in email messages using a combination of two-dimensional structural segmentation and one-dimensional grammatical constraints, integrated with weighted finite-state transducers. They assumed that the blocks had previously been separated, and used features based on the lay-out of email messages to identify signature blocks among all blocks in a message. The accuracy (90%) of their approach is not satisfactory and the efficiency is not high enough to reproduce the method on a large email collection.

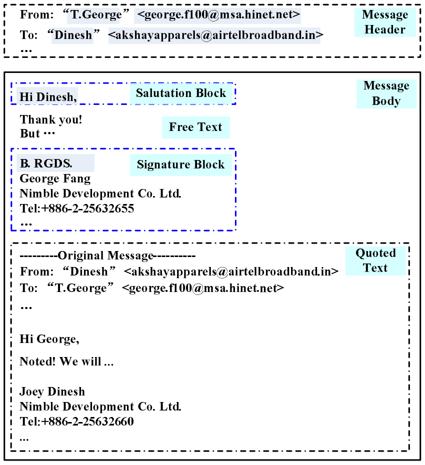
978-1-61284-181-6/11/$26.00 ©2011 IEEE

Carvalho et al.[9] have carried on an intensive research about the signature block locating, which used machine learning methods to detect and extract "signatures" in email messages that could serve as concise descriptions of the sender. They defined sets of signature feature patterns based on the content characteristic in signatures, and reduced the signature block detection and the signature lines extraction to classification problems. Although the maximal accuracy of Carvalho's report[9] is more than 97%, the efficiency isn't high when the methods are applied to a large email corpus. One reason is that the signature feature patterns are built on the content characteristic in signatures, and only after having analyzed the content of each line in the candidate signature block, can feature patterns matching process be performed, and these processes are very time-consuming. Another reason is that to extract the text in the signature block the methods must perform two steps of feature patterns matching: the first is detecting signature blocks and the second is extracting signature lines, which both reduce the performance more.

Since above researches suffer from precision in extracting salutation and signature lines or complexity in locating salutation and signature blocks, this paper is to propose a novel method to efficiently locate and extract salutation and signature blocks.

III. METHOD

A. Outline



From: "T.George" <george.f100@msa.hinet.net>
To: "Dinesh" <akshayapparels@airtelbroadband.in>
...

Hi Dinesh,

Thank you!
But ···

B. RGDS.
George Fang
Nimble Development Co. Ltd.
Tel:+886-2-25632655
...

---------Original Message----------
From: "Dinesh" <akshayapparels@airtelbroadband.in>
To: "T.George" <george.f100@msa.hinet.net>
...

Hi George,

Noted! We will ...

Joey Dinesh
Nimble Development Co. Ltd.
Tel:+886-2-25632660
...

Figure1. Example of a typical email message.

A typical email message is shown in Fig. 1. The message header comprises the basic metadata part of the email and consists of a set of RFC-822 header fields (from, to, cc, bcc, date, and subject). The rest of the message is considered the message body, which may include quoted text if the original was a reply or forward of another. The quoted text usually start with a system generated quoted tag that indicates the beginning of a quoted message, such as "original message", "forwarded by" or "forwarded message". The body of the quoted message may in turn include another quoted message and so on. We can further classify the lines of the message main body (excluding any quoted text) into salutation block, free text, and signature block. The salutation often appears in a separate line. The free text is considered the actual message that the sender intends to express to the recipients. The signature block may consist of a manually-typed signature (which we call a "free signature") and a relatively static set of system-generated signature lines.

In order to automatically locate and extract the salutation block and signature block from the email body, we must first split an email into the header and the body and then analyzing the Message body to locate the salutation line and the signature lines. And in the locating process, we must take into account the possibility that the body includes quoted text. The process flow of extracting salutation and signature blocks from an email message is shown in Fig. 2.

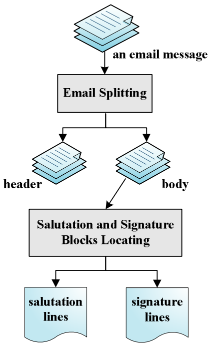


Figure2. Process flow of salutation and signature extracting.

B. The Steps of Locating Salutation and Signature Blocks

Conventionally, salutation lines appear at the beginning of the email body and consist of a word or phrase of greeting and a name alias of the recipient, such as "Dear X", "Hi, X", "Hello X" et al. And signature lines appear near the end of the email body and are composed of a word or phrase of cordiality and a name alias of the sender, such as "Sincerely Yours X", "Hon. X", "Yours X" et al. It is difficult to exactly extract salutation and signature blocks from email bodies, since the text in email bodies tends to be diverse, informal, and often creative. Carvalho et al.[9] used machine learning method to detect "signatures" in email messages and obtained high accuracy. But the machine learning method based on sets of signature feature patterns is very time-consuming when applied to the large email corpus. And it doesn't detect salutations and remove original message part in the end of email bodies in detecting signatures. For our system, we need to detect both signature blocks and salutation blocks. So we propose a novel block locating algorithm based on statistical and rules restricted methods, which exploit the statistical result and partial content characteristics of signature and salutation lines, and the results of our approach is efficient and relatively precise.

The steps of our Salutation and Signature Blocks Locating Algorithm based on statistical and rules restricted methods (shortened as SSBLA) are listed as follows.

1) *Step 1:* Analysing the body of an email and find out whether it includes the quoted text, and removing the corresponding part from the body.

If an email is a reply or forward of another, it will include a quoted text in the end of the body of this email. The quoted text may start with a system generated quoted tag that indicates the beginning of a quoted message, such as “original message”, “forwarded by” or “forwarded message”. So we must remove the quoted texts of an email message, and then can we carry out salutation and signature blocks detecting algorithm to the remainder.

2) *Step 2:* Exploiting a statistical method to roughly decide the lines that belong to salutation and signature blocks.

We locate the rough scope of salutation and signature blocks by selecting the first K1 lines of the body of every email message as the salutation block and the last K2 lines as the signature block. We make a statistical analysis on a great deal of email messages, and the result about the length distribution of salutation and signature blocks is shown in Table 1. The probability of salutation only including one line exceeds 60 percent, that is to say, most of salutation blocks in all of email messages include only the first line of the body. So we choose K1=1. To signature blocks, the probabilities that salutation blocks include 1, 2, and 3 lines are over 10 percent and as the length increases the frequency of corresponding signature blocks decreases. But when the length of signatures we select reaches 6 or 7, the frequency is close to 8 and even over 11. The reason is that there are usually not only the recipient’s name but also some other information about his contact, occupation or address in the end of email bodies, which makes the signature blocks reach 6 or 7 lines. So we choose K2=7 only if the total lines are no less than 9 lines in the whole email message (including the body text of at least one line, a blank line and a signature block of at most 7 lines), otherwise K2=3. In this way can we narrow the lines scope of extracting signature blocks and simultaneously promise a high precision.

Table 1. The length distribution of salutation and signature blocks.

	the proportion of different length(the number of lines) (%)										
	0	1	2	3	4	5	6	7	8	9	10
salutation block	34.04	61.70	4.26	0	0	0	0	0	0	0	0
signature block	9.91	19.80	23.76	10.89	3.96	4.95	11.88	7.92	0.99	1.98	2.90

3) *Step 3:* To correct the result of above-mentioned statistical method, introducing some restriction rules to decide the exact lines belonging to salutation and signature blocks.

According to the analysis to salutation and signature blocks in great quantity of actual emails, we can draw conclusions as follows.

a) For salutation blocks: If there is a salutation block in an email message, the length is usually one. So we introduce 3 rules to judge the cases that the salutation block does not exist and we neglect the salutation extracting step.

- *Rule 1:* The first line of the email body starts with “FYI” or “fyi”, which is the special instance that the signature block are at the beginning of email body.
- *Rule 2:* The first line ends with the punctuation describing the finish of a sentence, such as “?”, “!”, “.”.
- *Rule 3:* The salutation line usually consists of an alias of the recipient (at most 3 words) and a word or phrase conveying respect such as hello, my dear (at most 2 words). So if the number of words in the first line of the email body is more than 5, the salutation line does not exist.

b) For signature blocks: According to the above-mentioned statistical method, even if we know the signature block exist in an email, it is still difficult to decide the length of the signature. So we introduce 3 rules about the characteristic of the signature block to judge the cases that the salutation block does exist.

- *Rule 4:* If the first line of the email body starts with “FYI” or “fyi”, which means the following lines belong to signature block, we take the lines before the first blank line as the signature.
- *Rule 5:* Usually a normal signature block is separated from the main body by a blank line in an email message. So we can use the blank line to throw away the superfluous part in the signature block extracted via the above-mentioned statistical method.
- *Rule 6:* To a normal signature block, no matter how many lines a signature block includes, the words in each line of the signature block is not too many. And according to the analysis to a great deal of emails, the average length of each line in the signature block is not more than 36 characters.

C. *Algorithm*

According to above rules, it is easier to locate lines for salutation blocks than for signature blocks. As the signature blocks may appear in two different places in the email body: the beginning or the end, and the corresponding extracting methods are also different. We present the locating algorithms in the two circumstances as follows.

Formal definition:

- T is the plain text of the email body, and N is the total lines in T ;
- $r(t,i)$: the *ith* line in a text segment t , $i \geq 1$;
- $SalB$: the text segment of the extracted salutation block;
- $SigB$ is the text segment of the extracted signature block, and n is the total lines in $SigB$;
- $l(x)$ is the length of a text segment x , namely the amount of linguistically minimum elements in x , e.g. the linguistically minimum element of Chinese is

Chinese characters, and the linguistically minimum element of English is English word;

- ML_{SalB} : the maximum length of the salutation text line, e.g. ML_{SalB} =5 for in English emails;
- MAL_{SigB} : the maximum average number of characters in the signature text lines, e.g. MAL_{SigB} =36 for English emails;
- MBL : the maximum number of lines in the signature block, according to above analysis, $MBL = 7$.

The steps of our signature block locating algorithm from the beginning of the email body (shortened as SBLA_B) are shown in Fig. 3.

```
1.find out the first blank line in $T$ , and $bl$ is the serial number of the first blank line;
2.if( $bl == 0$ ) { $SigB = NULL$ , turn to the 6 step;};
3.if ( $bl > MBL + 1$ ) $bl = MBL + 1$ ;
4. $SigB$ is the top $n$ lines of $T$ ;
      $n = bl - 1$ , $SigB = \sum_{i=1}^{n} r(T,i)$ ;
5.calculate the average number of characters in $SigB$ . if it is over $MAL_{SigB}$ , the new $SigB$ is the top
  $m$ lines of the old $SigB$ , which satisfies the characters number restriction condition;
      if( $\frac{\sum_{i=1}^{n} l(r(SigB,i))}{n} > MAL_{SigB}$ )
      {   if( $\exists m, m = \max\{k \mid \frac{\sum_{i=1}^{k} l(r(SigB,i))}{k} \le MAL_{SigB}, 0 < k < n\}$ ) $SigB = \sum_{i=1}^{m} r(SigB,i)$ ;
          else $SigB = NULL$ ;
      }
6.output $SigB$ .
```

Figure 3. Steps of SBLA_B.

The steps of our signature block locating algorithm from the end of the email body (shortened as SBLA_E) are illustrated in Fig. 4.

```
1.if( $N \ge MBL + 1$ ) $n = MBL$ ; else $n = 3$ ;
2.find out the last blank line in $T$ , and $bl$ is the serial number of the last blank line;
3.if( $bl > 0 \&\& 0 < N - bl < n$ ) $n = N - bl$ ;
4. $SigB$ is the last $n$ lines of $T$ ;
      $SigB = \sum_{i=N-n+1}^{N} r(T,i)$ ;
5.calculate the average number of characters in $SigB$ , if it is over $MAL_{SigB}$ , the new $SigB$ is the last
  $n - m + 1$ lines of the old $SigB$ , which satisfies the characters number restriction condition;
      if( $\frac{\sum_{i=1}^{n} l(r(SigB,i))}{n} > MAL_{SigB}$ )
      {   if( $\exists m, m = \min\{k \mid \frac{\sum_{i=k}^{n} l(r(SigB,i))}{n-k+1} \le MAL_{SigB}, 0 < k < n\}$ ) $SigB = \sum_{i=m}^{n} r(SigB,i)$ ;
          else $SigB = NULL$ ;
      }
6.output $SigB$ .
```

Figure 4. Steps of SBLA_B.

The Salutation and Signature Blocks Locating Algorithm Based on Statistical and Rules Restricted Methods (shortened as SSBLA) is described in Fig. 5.

```
Algorithm：SSBLA

Input：the body $T$ of a plain-text email
Output：salutation text segment $SalB$ and signature text segment $SigB$
Begin：
1.find out the special line including “Original message”, “Forwarded by” or “Forwarded message” in $T$ ， and $sl$
  is the serial number of the found line;
2.if ( $sl > 0$ ) {delete $r(T,i)$ , $sl \le i < N$ }；
3.if ( $r(T,1)$ begins with “FYI” or “fyi”) { $SalB = NULL$ ， turn to step 6}；
4. if ( $r(T,1)$ ends with “?”、 “!” or “.”) { $SalB = NULL$ ， turn to step 8}；
5. if ( $l(r(T,1)) \le ML_{SalB}$ ) $SalB = r(T,1)$ ； else $SalB = NULL$ ；
6.call the procedure SBLA_B to locate the signature in the beginning of $T$ , $SigB$ =SBLA_B( );
7.if ( $SigB \ne NULL$ ) turn to the 8 step；
8.call the procedure SBLA_E to locate the signature in the end of $T$ , $SigB$ =TSBLA( );
9.output $SalB$ and $SigB$ .
End
```

Figure 5. Steps of algorithm SSBLA.

IV. EVALUATION

A. Datasets

In this section we analyze the methods described above. The experiments are carried out on the public Enron collection published by Federal Energy Regulatory Commission (FERC) in 2003[11]. It contains emails sent among 150 employees of Enron corporation from October, 1998 to June, 2002. A part of those emails include salutation and signature blocks with different kinds of format, and those emails can be used to evaluate the validity of the proposed blocks locating algorithm.

The emails in the collection are stored in folders, each folder correspond to one user and include several sub-folders such as sent_mail folder, inbox folder, all_documents folder and so on. In the experiments we select the sent_mail folders of 20 users, which include 6065 emails, and randomly choose 2000 emails from those folders. In the 2000 emails we selected randomly, after removing the quoted text from the email body，only 1672 emails have the text body, in which 358 emails include salutation blocks and 971 emails include signature blocks.

B. Experiments and Evaluation

First we analyze each email in the datasets and extract the To Header and the body of each email. Then we label the salutation and signature blocks manually. At last, we exploit above methods to automatically locate salutation and signature blocks and extract the corresponding text segments. To evaluate our method, we conduct the experiments on two locating methods. Method 1 is the statistical method illustrated in step 2 of section 3.3. Method 2 is the automatically locating algorithm based on statistical and rules restricted methods SSBLA proposed in this paper.

In the evaluation step, we compared the salutation and signature text segments extracted by the automatically locating methods with the corresponding manually labeled results. To evaluate the performance of the methods we use three measures: precision rate P, recall rate R and F1-measure F1, which are usually used to evaluate the performance in the Information Retrieval system. The formulas are defined as in (1):

$$P = n_{cor} \big/ n_{ext} \;;\quad R = n_{cor} \big/ n_{real} \;;\quad F1 = 2PR/(P+R) \qquad (1)$$

n_{cor} : the number of text blocks that agree with the labeled results in the extracted text blocks; n_{ext} : the number of the text blocks extracted by the locating method in the datasets; n_{real} : the number of the text blocks labeled manually in the datasets.

Precision rate and recall rate reflect two different aspects of locating performance, and the F1 measure reflects the integrated quality of locating methods. Table 2 shows the evaluation results of two methods in above datasets.

Table 2. evaluation results of two methods in above datasets.

		n_{real}	n_{ext}	n_{cor}	Precision (%)	Recall (%)	$F1$ measure (%)	Average $F1$ measure (%)
Method 1	salutation block	358	1672	339	20.28	94.69	33.41	18.52
	signature block	971	1672	48	2.87	4.94	3.62	
Method 2	salutation block	358	364	344	94.51	96.09	95.29	94.60
	signature block	971	985	922	93.60	94.95	93.91	

Table 2 shows that our approach is much better than the approach only based on the statistical method when applied to actual email datasets. The approach only based on the statistical method to locate the lines of salutation and signature blocks is so coarse that the locating performance is so low and F1 value dose not reach 20%. While both the precision rate and recall rate of our approach is above 93% by introducing several rules to refine the result of above-mentioned statistical method.

As we do not use any features of text lines or email lay-out features, the results of Chen et al.[10] and Carvalho et al.[9] can't be directly compared with ours. Our method is likely to be much more easily reproduced than forenamed methods, since it is based on both statistical method, which can improve the efficiency, and rules restricted method, which can promise the accuracy. Especially, as we build two key rules on the blank line and the average number of characters in signature lines to locate the signature lines in the end of email bodies, our approach have obtain a relatively high F1 measure (nearly 95%), which is higher than that of Chen et al.[10] (less than 92%) and close to that of Carvalho et al.[9] (96%). Moreover, our method can directly extract not only the signature lines but also the salutation line at the same time.

V. Conclusion and Future Work

In this work we addressed the problem of identifying salutation and signature blocks within the body of an email. We first split an email message into two parts: the header and the body. Then to the email body, we detect whether it includes the quoted text according to the system generated quoted tag and remove the corresponding part from the body and obtain the main body. The locating algorithm we proposed is applied to the main body and can simultaneously extract salutation and signature lines. We use the statistical method to roughly estimate the number of lines in salutation and signature blocks, and then introduce some restriction rules to refine the lines located by the statistical method. So the salutation and signature blocks locating algorithm can greatly improve the locating efficiency and at the same time promise a relatively high accuracy of the extracted blocks. Experiments on the public subset of the Enron corpus show that our approach achieves high accuracy (94%), recall (above 95%) and F1 measure (nearly 95%).

However, there are still some email instances, e.g. when there are some other texts before the salutation block or after the signature block in email bodies, the salutation or signature block can't be exactly located. These exceptional instances can be overcome by adding some new intensive rules or feature patterns, but how to keep the efficiency and the accuracy at the same time is still a difficult problem to research in the future.

References

[1] V. Bellotti, N. Ducheneaut, M. Howard and I. Smith. Taking email to task: the design and evaluation of a task management centered email tool. Proc. of the Conference on Human Factors in Computing Systems, Ft. Lauderdale, Florida.2003.

[2] H. Murakoshi, A. Shimazu and K. Ochimizu. Construction of Deliberation Structure in Email Communication. Pacific Association for Computational Linguistics, pp. 16-28, Waterloo, Canada, 1999.

[3] R. Sproat, J. Hu, and H. Chen. Emu: An e-mail preprocessor for text-to-speech. In 1998 Workshop on Multimedia Signal Processing, pages 239--244, Redondo Beach, CA, December 1998.

[4] C. Bird, A. Gourley and A. Swaminathan. Mining Email Social Networks. In: Proceedings of the 2006 international workshop on Mining software repositories. Shanghai, China, 2006: 137-143.

[5] C. Diehl, L. Getoor, and G. Namata. Name reference resolution in organizational email archives. In Proceddings of SIAM International Conference on Data Mining,Bethesda, MD , USA, April 20-22 2006.

[6] T. Elsayed, Oard D W. Modeling Identity in Archival Collections of Email. In: Proceedings of the Third Conference on Email and Anti-Spam. Mountain View, California, USA, 2006.

[7] T. Elsayed, D. W. Oard, and G. Namata. Resolving personal names in email using context expansion. In Association for Computational Linguistics(ACL), 2008.

[8] T. Elsayed, G. Namata, L. Getoor, and D. W. Oard. Personal name resolution in email: A heuristic approach. Technical Report UMIACS LAMP-TR-150, University of Maryland, March 2008.

[9] V. Carvalho and W. Cohen. Learning to extract signature and reply lines from email. In Proceedings of the 2004 Conference on Email and Anti-Spam (CEAS 04), August 2004.

[10] H. Chen, J. Hu, and R. Sproat. Integrating geometrical and linguistic analysis for e-mail signature block parsing. ACM Transactions on Information Systems, 17(4):343--366, October 1999.

[11] The email collection of Enron Corporation. http://www.cs.cmu.edu/~enron/. 2003.