基于邮件正文的邮箱用户别名抽取

尹美娟 陈庶民 刘晓楠 路 林
(信息工程大学信息工程学院 郑州 450002)

摘 要 邮箱用户身份信息挖掘是数据挖掘研究的一个热点。当前相关研究大多仅从邮件头中抽取邮箱用户的别名,遗漏了邮件正文中潜藏的更能代表通信双方身份的别名信息。针对纯文本邮件正文中邮箱用户别名信息抽取问题,提出了基于统计和规则过滤的称呼块和签名块定位算法,该算法能高效准确地从邮件正文中提取出蕴涵邮箱用户别名的称呼块和签名块文本片段;进一步提出了基于别名边界词汇模板修正的别名抽取方法,从而提高了仅基于命名实体识别或词性标注工具识别别名的准确率。实验结果表明,提出的方法可以有效地抽取出邮件正文中邮箱用户的别名。

关键词 实体解析,邮件正文,别名抽取,称呼块签名块定位,别名边界词汇模板
中图法分类号 TP391 **文献标识码** A

Extracting Name Aliases of Mailbox Users from Email Bodies

YIN Mei-juan CHEN Shu-min LIU Xiao-nan LU Lin
(Institute of Information Engineering, Information Engineering University, Zhengzhou 450002, China)

Abstract Mining user identity information from emails is an important research topic in data mining. Most approaches extract users' names only from the email headers, but names appearing in email bodies are usually more suitable for representing the sender's or recipient's identity. This paper focused on extracting users' name aliases in the body of plain-text emails. Firstly, to effectively elicit salutation and signature block from email bodies, a salutation and signature blocks locating algorithm based on statistical and rules restricted methods was proposed. Then to extract all valid aliases in the salutation and signature lines, a novel approach was proposed based on name boundary word template built on the characteristics of alias neighboring words, which can verify and amend aliases identified by named entity recognition or part-of-speech tagging tools. Results on Enron corpus indicate that the approaches proposed can efficiently and automatically extract user's aliases from email bodies.

Keywords Entity resolution, Email body, Alias Extraction, Salutation and signature blocks locating, Name boundary word template

1 引言

在虚拟化的网络世界中,人们进行网络通信时常使用不同形式的名字代替用户的真实姓名,如正式名、昵称、简称、敬称等(本文将其统称为别名),且习惯在一段时间内使用较为固定的别名。因此别名在一定程度上代表了用户在虚拟网络世界中的特定身份。从网络数据中挖掘用户的身份信息是目前数据挖掘研究的一个热点,该技术可应用于身份识别、信息检索、社会网络分析等应用研究领域。本文针对邮件通信数据中携带的用户身份信息,用邮件地址标识邮箱用户(不考虑一个用户拥有多个邮件地址的情况),研究邮箱用户的别名抽取方法。

现有相关研究大多从邮件头中抽取邮箱用户的别名,遗漏了邮件正文中潜藏的更能代表通信双方身份的别名。在邮件正文中,与代表邮箱用户的邮件地址直接关联的别名仅出现在称呼块与签名块中,故要抽取邮箱用户的别名,首先应定位并提取出邮件正文中的称呼块和签名块文本片段。为此,本文提出了基于统计和规则过滤的称呼块和签名块定位算法,该算法在提高块定位效率的同时保证了较高的准确率。在提取出邮件正文中的称呼块和签名块后,可直接利用命名实体识别工具或词性标注工具抽取潜在别名,但这往往存在召回不完整或无法召回的情况。为此,本文进一步提出了基于别名边界词汇模板修正的别名抽取算法,对识别出的潜在别名进行了校验和修正,提高了块内有效别名的识别率。在Enron邮件数据集上的实验结果表明,本文提出的块定位算法和别名抽取方法可有效地从邮件正文中抽取出邮箱用户的别名信息。

2 相关研究

邮箱用户的人名别名提取隶属实体解析范畴。实体解析通常是指将数据中出现的指代物(如名字、短语等)映射到现实世界中的实体(如人、地点等)的处理过程。目前实体解析

到稿日期:2011-05-19 返修日期:2011-08-18 本文受某国防基金资助。
尹美娟(1977—),女,博士生,讲师,主要研究方向为网络与信息安全、数据挖掘、社会网络等,E-mail:raindot_ymj@163.com。

的研究[1,2]已较成熟，但涉及到人名别名抽取的研究为数还不多。Bollegala D. 等[3-6]提出了利用训练出的模式从 Web 中抽取给定真实人名的候选别名的方法。该方法可以抽取出实体的别名，但无法将其与实体的邮件地址关联；另外，本文研究的是邮件数据集中人名别名的抽取问题，由于邮件正文书写的无格式性，很难用基于模式的方法来抽取邮件正文中的别名。

邮件中的人名指代解析和实体身份建模中均涉及邮箱用户的别名抽取。Bird C. 等[7]从邮件首部中抽取(别名，邮件地址)对，利用对间相似度进行聚类，解决用户的别名和多邮箱地址的关联问题。Diehl C. 等[8]首次提出邮件全文中提及的人名指代解析问题，其利用从邮件头中抽取出的邮件地址及其关联的人名，构建邮件通信社会网络，并通过分析邮件发送者之间的关系紧密程度，来解析人名指代。由于从邮件正文中提取邮箱用户的别名难度较大，上述两个文献均未涉及邮件正文中邮箱用户的别名抽取问题。Elsayed T. 等[9-11]研究的也是邮件正文中的人名指代解析问题，其将邮件地址作为描述实体身份的关键属性，从邮件头、邮件正文中的称呼和签名部分抽取邮件地址关联的人名，从而构建实体身份模型以解析人名指代。其定位签名块的方法是利用空行来划分正文内容与签名块的界限，该方法对格式较统一的 Enron 邮件比较有效，但面对无统一的标准格式的普通邮件数据时，定位不准确；在从称呼块和签名块中提取别名时，利用一些简单的规则去除一些停用词或无效的句子后，将块中保留下来的文本行与邮件地址中@前的用户名进行字符串比较，若有相似的词，则将整个文本行作为该邮件地址用户的别名。显然，该方法提取出的别名的准确率和召回率均有待提高。

在邮件正文称呼块签名块定位方面，少数研究利用邮件消息的布局或机器学习的方法来定位。Chen H. 等[12]使用结合了二维结构划分与一维语法约束的带权有限状态转换器方法来分析邮件消息中的签名块域。由于算法复杂度较高，应用于较大邮件数据集时，其准确率仅达 90% 且效率也不够高。Carvalho V. 等[13]利用机器学习方法检测并抽取出邮件消息中简述发件人信息的签名行，根据签名行的内容特征定义签名块特征模式，从而将签名块检测和签名行文本抽取转化为分类问题。该方法可以达到高于 97% 的准确率，但由于分析签名行文本内容和特征模式匹配均很耗时，算法效率较低。

在应用于大型邮件数据集时，上述方法在称呼块签名块定位或块内别名识别及抽取方面还存在准确率不够高或效率较低的问题。针对邮件正文中邮箱用户别名抽取的两个关键问题：称呼块签名块定位、块内别名抽取，本文提出新的方法，以提高邮件正文中邮箱用户别名自动抽取的准确性和效率。

3 称呼块和签名块定位

如图 1 所示，邮件消息中可直接与邮箱用户关联的别名信息出现在邮件头和邮件正文中。邮件头的各地址字段，如“From”、“To”、“Cc”和“Bcc”等，直接给出了邮箱用户的别名和邮件地址；正文开头称呼语中的人名是邮件头中“To”字段标记的邮箱用户的别名，正文签名块中的落款名是邮件头中“From”字段标记的邮箱用户的别名。

为实现邮件正文中邮箱用户的别名抽取，必须解决邮件正文中称呼块和签名块文本片段的有效定位问题。为此，本文提出了基于统计和规则过滤的块定位方法。

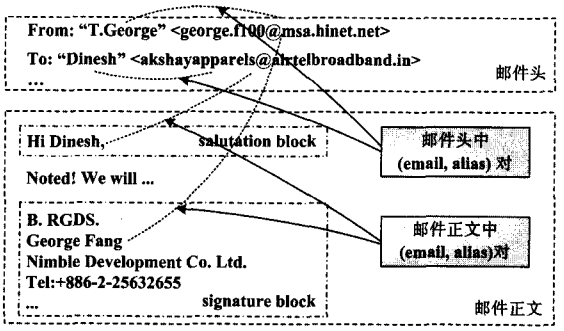

图 1 邮件数据实例

3.1 称呼块签名块定位步骤

基于统计和规则过滤的块定位的基本步骤如下。

第一步 分析邮件正文中是否含有被回复邮件的内容，若有，则去除此部分内容，得到有效邮件正文。

很多情况下，用户发送的邮件是对收到的邮件的回复。在生成回复邮件时，邮件系统往往会自动在邮件正文后面追加原始邮件内容。因此在提取邮件签名块前，必须去除由 original message、forwarded by 或 forwarded message 等标注的自动追加部分。

第二步 利用统计方法，对称呼块和签名块包含的文本行数进行大致定位。

取邮件正文的开头 K_1 行作为称呼块、最后 K_2 行作为签名块，从而大致定位称呼块和签名块包括的文本范围。通过对大量邮件数据的统计，不同长度(即行数)称呼块和签名块比例的分布如表 1 所列。

表 1 不同长度称呼块和签名块比例的分布

	长度(%)										
	0	1	2	3	4	5	6	7	8	9	10
称呼语	34.04	61.70	4.26	0	0	0	0	0	0	0	0
签名块	9.91	19.80	23.76	10.89	3.96	4.95	11.88	7.92	0.99	1.98	2.90

可见，$K_1=1$ 的称呼块出现频率最大，凡是存在称呼块的邮件，绝大部分长度为 1 行，因此，本文选取 $K_1=1$。对签名块而言，$K_2=1,2,3$ 的签名块出现概率较大，随着长度增加出现频率明显减小；$K_2=6、7$ 的签名块出现频率也较大的原因是，很多邮件在签名块中除了给出发件人名字外，还含有发件人的其他联系信息。因此，为了尽可能缩小抽取范围，并保证提取准确度，本文采取的策略是：若邮件正文总行数大于等于 9(除了签名块之外，必然还含有至少 1 行的正文内容及一个空行)，则选取 $K_2=7$；否则，取 $K_2=3$。

第三步 引入块判定规则对上述基于统计方法得到的块文本行进行过滤。

根据上述统计结果可知，若邮件中有称呼块，则通常为正文的第一行。本文引入以下规则排除首行不是称呼块的情况。

规则 1 正文第一行文本以“FYI”或“fyi”开头，或以“?”，“!”，“.”结束。

正文第一行文本以“FYI”或“fyi”开头，表示此后紧随的是关于发件人的信息，即签名信息；正文第一行文本以“?”，“!”，“.”结束，表示这是一个句子，而不是称呼块。

规则 2 收件人地址大于 1。

若收件人地址个数大于 1，则不能确定与签名块中别名

配对的邮件地址。

规则 3　称呼行文本长度大于阈值。

根据上述统计方法可知，签名块的长度及其存在性较难确定。本文引入以下规则，提高签名块定位的准确度。

规则 4　正文第一行以“FYI”或“fyi”开头，从正文头部提取签名块。

若正文第一行以“FYI”或“fyi”开头，则说明其后为签名信息，故忽略正文尾部签名块的提取，而提取正文中第一个空行前的内容作为签名块。

规则 5　在签名块允许的长度范围内，利用空行来限定提取对象。

通常较正规的签名块，无论其长度多大，都会用一个空行与正文内容分隔开。

规则 6　签名块内每行文本的平均长度小于阈值。

3.2　基于统计和规则过滤的块定位算法

相关定义：

T：一封邮件的邮件正文文本，N：T 中文本总行数；

$r(t,i)$：文本段 t 中第 i 行文本，$i \geqslant 1$；

$SalB$：抽取出的称呼块文本片段 Salutation Block；

$SigB$：抽取出的签名块文本片段 Signature Block；

$l(x)$：文本片段 x 的长度，即 x 中包含的语言文字最小语素个数，如中文最小语素为汉字、英文最小语素为单词；$l(SalB)$、$l(SigB)$分别表示称呼块和签名块文本片段的长度；

ML_{SalB}：称呼块文本行的最大长度，例如，英文 $ML_{SalB}=5$；

MAL_{SigB}：签名块内各行文本平均长度的最大值，例如，英文 $MAL_{SigB}=36$；

MBL：签名块内文本最大行数，根据上述统计分析，$MBL=7$。

基于上述定义，邮件正文首部的签名块提取算法（记作 SBLA_B）的基本步骤如图 2 所示。

1. 在 T 中查找第一个空行的位置，记其所在行号为 bl；
2. if($bl==0$) {$SigB=NULL$，转 6}；
3. if ($bl>MBL+1$) $bl=MBL+1$；
4. 取 T 的前 n 行为签名块 $SigB$；

$$n=bl-1,\quad SigB=\sum_{i=1}^{n} r(T,i);$$

5. 判断 $SigB$ 中各行文本的平均长度，若大于 MAL_{SigB}，则仅提取满足该约束条件的前 m 行文本：

$$\text{if}\left(\frac{\sum_{i=1}^{n} l(r(SigB,i))}{n} > MAL_{SigB}\right)$$

$$\{\ \text{if}\left(\exists m,\ m=\max\left\{k \mid \frac{\sum_{i=1}^{k} l(r(SigB,i))}{k} \leq MAL_{SigB}, 0<k<n\right\}\right)\ SigB=\sum_{i=1}^{m} r(SigB,i);$$

else $SigB=NULL$；

}

6. 输出 $SigB$。

图 2　邮件正文首部的签名块提取步骤

1. if($N \geq MBL+1$) $n=MBL$；else $n=3$；
2. 在 T 中查找最后一个空行的位置，记其所在行号为 bl；
3. if($bl>0$&&$0<N-bl<n$) $n=N-bl$；
4. 取 T 的后 n 行为签名块 $SigB$；

$$SigB=\sum_{i=N-n+1}^{N} r(T,i);$$

5. 判断 $SigB$ 中各行文本的平均长度，若大于 MAL_{SigB}，则仅提取满足该约束条件的后 $n-m+1$ 行文本：

$$\text{if}\left(\frac{\sum_{i=1}^{n} l(r(SigB,i))}{n} > MAL_{SigB}\right)$$

$$\{\ \text{if}\left(\exists m,\ m=\min\left\{k \mid \frac{\sum_{i=m}^{n} l(r(SigB,i))}{n-k+1} \leq MAL_{SigB}, 0<k<n\right\}\right)\ SigB=\sum_{i=m}^{n} r(SigB,i);$$

else $SigB=NULL$；

}

6. 输出 $SigB$。

图 3　邮件正文尾部的签名块提取步骤

基于上述定义，邮件正文尾部的签名块提取算法（记作 SBLA_E）的基本步骤如图 3 所示。

基于上述两个子过程，邮件正文中称呼块签名块定位算法如图 4 所示。

算法：基于统计和规则过滤的称呼块签名块定位算法 SSBLA

Input：一封纯文本邮件的邮件正文文本 T、邮件头 To 字段中的地址个数 nTo

Output：称呼块文本片段 $SalB$ 和签名块文本片段 $SigB$

Begin:

1. 在 T 中查找含“Original message”、“Forwarded by”或“Forwarded message”的特定文本行，记其所在行号为 sl；
2. if ($sl>0$) {delete $r(T,i)$, $sl \leq i < N$)；
3. if ($r(T,1)$以“FYI” or “fyi”开头) {$SalB=NULL$，转 6}；
4. if ($r(T,1)$以“?”、“!” or “.”结束)||($nTo>1$)) {$SalB=NULL$，转 7}；
5. if ($l(r(T,1)) \leq ML_{SalB}$) $SalB=r(T,1)$；else $SalB=NULL$；
6. $SigB=SBLA\_B()$；if ($SigB \neq NULL$) 转 8；
7. $SigB=SBLA\_E()$；
8. 输出 $SalB$ 和 $SigB$。

End

图 4　基于统计和规则过滤的块定位算法

4　块内邮箱用户别名抽取

在提取出邮件正文中称呼块和签名块文本片段后，对文本片段进行词性标注或命名实体识别，利用标注结果识别潜在别名。然而，邮件正文中出现的邮箱用户的别名很多情况下是昵称、简称、尊称、敬称等非正式人名，导致依赖命名实体识别或词性标注工具得到的潜在别名，可能存在召回不完整或无法召回的情况。为此，本文利用邮件正文中称呼块和签名块的行文特征构造别名边界词汇模板，对经词性标注或命名实体识别工具标注出的潜在别名进行修正，以提高邮箱用户别名抽取的准确度。

4.1　别名边界词汇模板的定义

（1）确定别名的长度

不同语言文字中的人名均有两类形式：正式名和非正式名，如昵称、简称、尊称、敬称等。以英文人名为例，正式名通常由 2 个单词或 3 个单词（first name、middle name 和 last name）组成，非正式名通常由 1 个单词（first name 或 last name）或 2 个单词（first name 或 last name，以及表示昵称、简称、尊称、敬称或职位等的称呼语单词）组成。根据上述分析，定义潜在别名的长度、最小长度及最大长度。

定义 1（潜在别名的长度、最小长度及最大长度）

$l(x)$：词语序列 x 的长度，即 x 中包含的语言文字最小语素个数，如中文最小语素为汉字、英文最小语素为单词。

若 x 为标点符号，则 $l(x)=1$；若 $l(x)$为 0，则表示词语序列 x 不存在；若 n 表示潜在人名词串，则 n 的长度是 $l(n)$。

$L_{n\min}$：潜在人名词串 n 可能的最小长度。

$L_{n\max}$：潜在人名词串 n 可能的最大长度。

若 n 为中文潜在人名，则 $L_{n\max}=4$、$L_{n\min}=1$；若 n 为英文人名，则 $L_{n\max}=3$、$L_{n\min}=1$。

（2）构建别名前后界词汇表

根据邮件正文中称呼块或签名块中别名前后的特殊上下文词语，构建别名前、后界词汇表。按照书信习惯，称呼语通常由用于表达问候的词或短语及收件人的称呼（可以是正式名、尊称、敬称、昵称等）组成，如“亲爱的 XX”，“Dear XX”等；签名块中通常包含发件人的落款署名，如“联系人：XX”、“Yours sincerely XX”等。通过对邮件语料库中称呼块和签名块中的文本进行统计，得到邮件称呼语和签名块中别名前、后常见的专有词汇，表 2 列出了中英文邮件中出现频度较高

的别名前后界词汇。

表 2　称呼块签名块中人名前后界词汇示例

英文	前界	Dear, My dear, Hi, "Hi,", Hello, "Hello,", Honorable, "Hon.", Yours, Yours sincerely, Sincerely yours, Sincerely, Yours faithfully, Faithfully yours, Yours truly, Truly yours, Yours respectfully, Respectfully yours, cordially
	后界	',', SPACE, CRLF
中文	前界	亲爱的，敬爱的，尊敬的
	后界	',', SPACE, CRLF, ':', 你好，您好，敬上

定义 2(前、后界词的长度、最大长度)

f, r：分别表示前、后界词，则 f, r 的长度为 $l(f)$, $l(r)$；当 f 或 r 为标点符号或特殊字符，如 SPACE、CRLF，则 $l(f)=1$；

$L_{f\max}$：前界词最大长度；

$L_{r\max}$：后界词最大长度。

不同语言的别名其前、后界词的最大长度不同，中文人名：$L_{f\max}=3$、$L_{r\max}=2$；英文人名：$L_{f\max}=2$、$L_{r\max}=1$。

(3) 定义邮件称呼块签名块中别名边界词汇模板

定义 3(别名边界词汇模板)

在邮件正文称呼块或签名块中，若存在词语序列〈fnr〉，满足 f 是前界词、r 是后界词、n 为标注出的潜在别名或 n 中的部分为标注出的潜在别名，且 $0<=l(f)<=L_{f\max}$、$0<=l(r)<=L_{r\max}$、$l(f)\times l(r)\neq 0$、$L_{n\min}<=l(n)<=L_{n\max}$，则称〈$fnr$〉为邮件正文中邮箱用户的别名边界词汇模板，记为 $FNR1$。

当潜在别名未被命名实体识别或词性标注工具识别出来时，通过上述模板无法抽取出此类潜在别名。为此，需要定义更严格的别名边界词汇模板。定义如下：

邮件正文称呼块或签名块中，若存在词语序列〈fnr〉，满足 f 是前界词、r 是后界词、n 为任意词语序列，且 $0<l(f)<=L_{f\max}$、$0<l(r)<=L_{r\max}$、$L_{n\min}<=l(n)<=L_{n\max}$，则称〈$fnr$〉为邮件正文中邮箱用户的别名边界词汇模板，记为 $FNR2$。

4.2　基于别名边界词汇模板修正的别名抽取算法

基于别名边界词汇模板的别名抽取算法的基本思想是：若块内文本中有被标注为人名的潜在别名，依据别名边界词汇模板 $FNR1$ 确定其前后边界，对潜在别名的首部和尾部进行修正，修正后得到的潜在别名 n 即为待抽取的别名；否则，根据模板 $FNR2$ 查找可确定前后边界的词语序列 n，则找到的词语序列 n 就是待抽取的别名。

定义：

T：经词性标注或命名实体识别后的邮件正文称呼块或签名块文本片段；

$w(i)$：T 中第 i 个语言文字最小语素(英文为单词、中文为汉字)；

$w(i)..w(j)$：表示 T 中的一个词语序列；

n：在 T 中被标注为人名的词语序列 n；

x：n 在 T 中的词序号，即词语序列 n 中第一个单词(或文字)在 T 中的序号。

基于上述定义，依据 $FNR1$ 修正潜在别名 n 的子过程(记作 AEA_FNR1)的基本步骤如图 5 所示。

依据 $FNR2$ 抽取潜在别名 n 的算法(记作 FNR2A)的基本步骤如图 6 所示。

```
Procedure AEA_FNR1(Table 1, T, n, x)
if (l(n) ≥ L_nmax) return n;
if ((x+l(n) ≤ a < l(T)) && (0 ≤ b < L_rmax) && (a+b < l(T)) && (w(a)..w(a+b) 是表 1 中的后界词))
{
    temp ← w(x)..w(a−1);
    if (l(temp) <= L_nmax) n ← temp;
    if (l(n) == L_nmax) return n;
}
if ((0 ≤ a < x) && (0 ≤ b < L_fmax) && (a+b < x) && (w(a)..w(a+b) 是表 1 中的前界词))
{
    temp ← w(a+b+1)..w(x+l(n)−1);
    if (l(temp) <= L_nmax) n ← temp;
}
return n;
```

图 5　基于模板 FNR1 的别名修正子程序 AEA_FNR1

```
Procedure AEA_FNR2(Table 1, T)
n ← NULL;
if ((0 ≤ i < l(T)) && (0 ≤ a < L_fmax) && (i+a < l(T)) && (w(i)..w(i+a) 是表 1 中的前界词))
    if ((i+a < j < l(T)) && (0 ≤ b < L_rmax) && (j+b < l(T)) && (w(j)..w(j+b) 是表 1 中的后界词))
        if (j−(i+a) > 1)
        {   temp ← w(i+a+1)..w(j−1);
            if (l(temp) <= L_nmax) n ← temp;
        }
return n;
```

图 6　基于 FNR2 的别名修正子程序 AEA_FNR2

基于别名边界词汇模板修正的别名抽取算法(记作 NBWT_AEA)的步骤描述如图 7 所示。

```
Procedure NBWT_AEA(Table 1, T)
n ← NULL;
if ((0 ≤ x < l(T)) && (j ≥ 0) && (x+j < l(T)) && (w(x)..w(x+j) 是一个标注出的人名))
{
    n ← w(x)..w(x+j);
    if (l(n) < L_nmax) n ← AEA_FNR1(Table1, T, n, x);
}
else
    n ← AEA_FNR2(Table1, T);
return n;
```

图 7　基于别名边界词汇模板修正的别名抽取算法 NBWT_AEA

该算法具有较好的普适性，只要将上述定义中的边界词汇列表、语言文字的最小语素单位及相关数值进行适当设定，即可适用于不同语种的邮件正文别名抽取。本文在实验部分仅针对英文邮件进行了实验，效果较好。

5　实验设计与结果分析

实验采用 FERC (Federal Energy Regulatory Commission)在 2003 年公开的 Enron 邮件公共语料[15]为测试数据集，其包含了 Enron 公司 1998 年 10 月到 2002 年 6 月共 150 名员工间相互通信的邮件。其中很多邮件在正文中包含了形式各异的称呼块和签名块，可用来测试本文提出的块定位算法的有效性；此外，该邮件语料中还提供了一个邮件地址与用户姓名的映射表，可为本文的基于人名边界词汇模板的别名抽取算法实验结果提供评价依据。

语料中的邮件集以每个用户为单位的文件夹形式存储，其中包括发件箱、收件箱等多个子文件夹。实验选取其中 20 个人的 sent_mail 子文件夹中共 6065 封邮件，从中随机选出

2000 封包含数据集提供的“邮箱地址-人名”映射表中相应人名的邮件。在 2000 封测试邮件中，去除被回复邮件的相应部分后，有 1672 封有正文，其中 348 封有称呼块、971 封有签名块。

5.1 称呼块和签名块定位实验

首先对测试数据集中的每封邮件进行解析，提取出邮件首部的地址字段和邮件正文；然后定位邮件正文中的称呼块和签名块范围，并提取出相应的文本片段。实验采用两种定位方法测试算法性能，方法 1：直接利用统计方法，大致定位称呼块和签名块的文本行数，进而提取称呼块和签名块；方法 2：利用本文提出的基于统计和规则过滤的定位方法，提取邮件的称呼块和签名块。

用衡量信息抽取系统性能常用的 3 个指标，即准确率 P、召回率 R 和 F-measure 来评价定位算法的有效性。就本实验而言，准确率 P 是定位得到的文本块中与人工标注一致的文本块个数占定位得到的文本块总个数的比率，召回率 R 是定位得到的文本块中与人工标注一致的文本块个数占人工标注出的文本块总个数的比率，其数学公式表示如下：

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

式中，n_{cor} 表示定位得到的文本块中与人工标注一致的文本块个数，n_{ext} 表示定位得到的文本块总个数，n_{real} 表示测试语料中实际标注出的文本块个数。P 值反映了块定位的准确程度，R 值反映了块定位的完备程度，$F1$ 值反映了块定位的综合质量，$F1$ 值越大，方法的性能越好。

对方法 1 和方法 2 计算上述 3 种评价指标，测试结果如表 3 所列。

表 3 两种块定位方法的实验结果

测试类型		实际标注出的块数	定位得到的块数	与标注一致的块数	准确率(%)	召回率(%)	F1 值(%)	平均 F1 值(%)
方法 1	称呼块	358	1672	339	20.28	94.69	33.41	18.52
	签名块	971	1672	48	2.87	4.94	3.62	
方法 2	称呼块	358	364	344	94.51	96.09	95.29	94.60
	签名块	971	985	922	93.60	94.95	93.91	

实验结果表明，在面向实际邮件数据时，仅基于统计结果定位称呼块和签名块文本行数的方法过于粗糙，导致准确率不高、召回率非常低，平均 F 值还不到 20%。而引入了块判定规则后，可以较准确地定位称呼块和签名块的具体文本范围，基于统计和规则过滤的块定位算法准确率和召回率均达到了 93%以上。但由于邮件书写具有随意性，仍存在方法 2 无法准确定位的情况，例如，当邮件第一行文本较短但并非称呼块时，会被误认为称呼块；少数邮件其称呼块长度大于 2 行或在称呼块前有其他内容时，不能正确抽取；当邮件末尾的文本特征与签名块特征相似，但并非签名块时，会被误认为签名块；少数邮件在签名块之后还有一些其他正文文本，导致无法正确抽取签名块。

由于本文的块定位算法未使用任何称呼行签名行文本特征或邮件布局特征，因此，不能与 Chen[12] 和 Carvalho[13] 的实验结果进行直接比较。但相比之下，由于本文的块定位算法在利用统计方法提高算法效率的同时，利用规则过滤保证了算法的准确性，因此，本文的方法具有更好的可扩展性。尤其是，引入了空行和行文本平均字符数这两个过滤规则，使得本文方法可以得到接近 95% 的 $F1$ 值，该值高于 Chen 的方法（不到 92%）且接近 Carvalho 的方法（96%）。此外，本文提出的块定位算法可同时抽取邮件正文中的签名行和称呼行。

5.2 块内邮箱用户别名抽取实验

将实验一中人工标注出的 358 个称呼块和 971 个签名块文本片段（均为英文）作为本实验的测试数据，首先从中抽取出有效别名，分别与所在邮件的邮件头中相应用户的邮件地址关联，得到别名信息对（email，alias）；然后验证邮箱用户别名抽取算法的有效性。采用 3 种方法抽取块内别名，方法 1 直接利用英文领域知名的命名实体识别工具——斯坦福大学的 Named Entity Recognizer Systerm Version1. 1. 1[14]，对块内人名进行标注，提取标注出的人名作为邮箱用户关联的别名；方法 2 利用基于别名边界词汇模板 $FNR1$ 的别名抽取子过程 AEA_FNR1 抽取块内别名；方法 3 利用基于别名边界词汇模板 $FNR1$ 和 $FNR2$ 的别名抽取算法 NBWT_AEA 抽取块内别名。

抽取出的邮箱用户的别名字符串，若与人工标注的结果一致，则认为该别名正确。与实验一相似，仍用衡量信息抽取系统性能的准确率 P、召回率 R 和 F-measure 来评价别名抽取算法的有效性，其数学公式表示如下：

$$P=n_{cname}/n_{ename}\ ,R=n_{cname}/n_{mame}\ ,F1=2PR/(P+R)$$

式中，n_{cname} 表示成功抽取出的别名中正确的别名数，n_{ename} 表示成功抽取出的别名总个数，n_{mame} 表示测试数据中实际标注出的别名个数。

对方法 1 和方法 2 计算上述 3 种评价指标，测试结果如表 4 所列。

表 4 3 种别名抽取方法的实验结果

	实际标注的别名数	抽取出的别名个数	正确抽取的别名个数	准确率(%)	召回率(%)	F1 值(%)
方法 1	1287	698	672	96.28	52.29	67.77
方法 2	1287	698	686	98.28	53.30	69.12
方法 3	1287	1079	1053	97.59	81.82	89.01

与方法 1 相比，方法 2 也无法抽取出被 Stanford NER 工具遗漏的别名，故其抽取出的别名数与方法 1 相同，但由于引入模板 $FNR1$ 修正了 14 个不完整的人名，使得其准确率有所提高。但对于不满足模板 $FNR1$ 的不完整人名，方法 2 亦束手无策。

经分析，由于 Stanford NER 工具自身的误差，导致实验数据中一部分人名无法召回，例如，Phillip、Theresa 等人名被识别为地名，Darrell、Shelley 等人名被识别为组织名。然而，这些人名中很多满足模板 $FNR2$，可通过方法 3 准确识别和抽取，故方法 3 的召回率达到了 81%。在准确率方面，由于基于模板 $FNR2$ 抽取的别名其准确度低于基于模板 $FNR1$ 识别出的别名，因此方法 3 的准确度略低于方法 2。从 $F1$ 值来看，方法 3 的综合性能要远高于方法 2 和方法 1。

结束语 针对目前多数研究仅从邮件头中抽取别名，造成信息利用不充分、别名信息不全面的问题，本文提出从邮件正文称呼块和签名块中提取邮箱用户别名信息的思想。为实现邮件正文中称呼块和签名块文本片段的准确定位和提取，提出了基于统计和规则过滤的块定位算法；在利用命名实体识别或词性标注工具标注出称呼块签名块文本片段中的潜在别名后，提出了基于别名边界词汇模板修正的别名抽取算法，

（下转第 199 页）

参 考 文 献

[1] 陈康,郑纬民. 云计算:系统实例与研究现状[J]. 软件学报,2009,20(5):1337-1348

[2] Buyya R, Yeo C S. Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility[J]. Future Generation Computer Systems, 2009(25): 599-616

[3] Lee Y, Wang C. Profit-driven service request scheduling in clouds[C]// IEEE/ACM International Conference on Cluster, Cloud and Grid Computing. 2010: 15-24

[4] Stober J, Neumann D. Market-based pricing in grids: on strategic manipulation and computational cost[J]. European Journal of Operational Research, 2010, 203: 464-475

[5] Mihailescu M, Teo Y M. On economic and computational-efficient resource pricing in large distributed systems[C]// IEEE/ACM International Conference on Cluster, Cloud and Grid Computing. 2010: 838-843

[6] Saure D, Sheopuri A. Time-of-use pricing policies for offering cloud computing as a service[C]// IEEE International Conference on Service Operations and Logistics and Informatics. 2010: 300-305

[7] Mihailescu M, Teo Y M. Dynamic Resource Pricing on Federated Clouds[C]// IEEE/ACM International Conference on Cluster, Cloud and Grid Computing. 2010: 513-517

[8] Paleologo G A. Price-at-Risk: A methodology for pricing utility computing services[J]. IBM Systems Journal, 2004, 43(1): 20-31

[9] Chen Y, Das A. Pricing-based strategies for autonomic control of web servers for time-varying request arrivals[J]. Engineering Applications of Artificial Intelligence, 2004(17): 841-854

[10] Ouyang J, Sahai A. A mechanism of specifying and determining pricing in utility computing environments[C]// IEEE/IFIP International Workshop on Business-Driven IT Management. 2007: 39-44

[11] Buyya R, Yeo C S. Market-oriented cloud computing: vision, hype, and reality for delivering IT services as computing utilities [C]// 10th IEEE International Conference on High Performance Computing and Communications. 2008: 5-13

[12] Henzinger T A, Singh A V. FlexPRICE: Flexible provisioning of resources in a cloud environment[C]// IEEE International Conference on Cloud Computing. 2010: 83-90

[13] Yeo C S, Venugopal S. Autonomic metered pricing for a utility computing service[J]. Future generation computer systems, 2010(26): 1368-1380

[14] Woitaszek M, Tufo H M. Developing a cloud computing charging model for high-performance computing resources[C]// IEEE International Conference on Computer and Information Technology. 2010: 210-217

[15] Jiang G, Cybenko G. Functional validation in grid computing [J]. Autonomous Agents and Multi-agent Systems, 2004(8): 119-130

[16] Chandra A, Gong W, Shenoy P. Dynamic resource allocation for shared data centers using online measurements[C]// ACM/IEEE International Workshop on Quality of Service. 2003: 381-400

[17] Furht B, Escalante A. Handbook of cloud computing[Z]. New York: Springer, 2010

(上接第 186 页)

利用该算法对潜在别名进行校验和修正,进而抽取出块内所有有效别名。在 Enron 邮件数据集上的实验结果表明,本文提出的基于统计和规则过滤的块定位算法可较为准确、完整地提取出邮件正文中的称呼块和签名块文本片段;同时,别名边界词汇模板的引入,在很大程度上提高了仅依据命名实体识别工具提取出的别名的有效性和完整性。

参 考 文 献

[1] Sarawagi S, Bhamidipaty A. Interactive deduplication using active learning[C]// ACM Special Interest Group on Knowledge Discovery and Data Mining. 2002

[2] Bhattacharya I, Getoor L. A latent dirichlet model for unsupervised entity resolution[C]// The SIAM International Conference on Data Mining (SIAM-SDM). Bethesda, MD, USA, 2006

[3] Bollegala D, Matsuo Y, Ishizuka M. Disambiguating personal names on the web using automatically extracted key phrases[C]// Proc. of the 17th European Conference on Artificial Intelligence. 2006: 553-557

[4] Bollegara D, Matsuo Y, Ishizuka M. Extracting key phrases to disambiguate personal names on the web[C]// Proc. CICLing 2006. 2006

[5] Bollegala D, Honma T. Identification of Personal Name Aliases on the Web[C]// Proceedings of WWW 2008 Workshop on Social Web Search and Mining (SWSM 2008). Beijing, China, 2008

[6] Bollegala D, Honma T, Matsuo Y, et al. Mining for personal name aliases on the web[C]// Proceeding of the 17th international conference on World Wide Web. Beijing, China, April 2008

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

[8] Diehl C, Getoor L, Namata G. Name reference resolution in organizational email archives[C]// Proceddings of SIAM International Conference on Data Mining. Bethesda, MD, USA, April 2006

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

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

[11] Elsayed T, Namata G, Getoor L, et al. Oard. Personal name resolution in email: A heuristic approach[R]. UMIACS LAMP-TR-150. University of Maryland, March 2008

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

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

[14] Stanford University. Named Entity Recognition System [EB/OL]. http://nlp.stanford.edu/software/stanford-ner-2009-01-16.tgz, 2009

[15] The email collection of Enron Corporation [DB/OL]. http://www.cs.cmu.edu/~enron/, 2003