

TEXEM : 一种基于实体的邮件任务提取策略

张相於¹ 陈继东² 李玉坤¹ 孟小峰¹

¹ (中国人民大学信息学院, 北京 100872)

² (EMC中国实验室, 北京 100084)

(zhangxy@live.com)

摘要 在信息化飞速发展的今天, 电子邮件的使用正在变得越来越频繁, 而且其应用场合也在不断扩展, 目前世界上很大一部分、并且越来越多的商业和个人往来都是通过电子邮件完成的。电子邮件的作用正在被人们不断拓展, 特别是它作为任务管理和协作管理的重要工具和手段, 被人们广泛使用。但是由于电子邮件数量越来越大, 其设计初衷只是单纯的通讯手段, 具有轻量性和随意性等特点, 因此当前的邮件管理工具很难有效的组织和管理人们的大量邮件信息, 更重要的是它们不能很好的完成邮件任务管理等扩展任务。本文提出了一种基于实体的邮件任务提取框架——TEXEM, 充分考虑电子邮件中的结构信息, 将邮件通过实体聚类的方法转化为任务的集合, 并对用户任务重要性进行评估, 使用户不再单纯依靠毫无结构的纯文本内容来处理邮件, 以起到辅助用户邮件处理过程、提高邮件管理效率的作用。

关键词 任务提取; 实体识别; 邮件处理; 聚类

中图法分类号 TP393

TEXEM : An Entity-based Task Extraction Approach for Emails

Zhang Xiangyu¹, Chen Jidong², Li Yukun¹, Meng Xiaofeng¹

¹(School of Information, Renmin University of China, Beijing 100872)

²(EMC Research China, Beijing 100084)

Abstract With the development of information technology, usage of email is getting more and more popular and intense. Usage of email has been extended to many scenarios. Nowadays, a large portion of commercial and personal communication is delivered through emails, and the number never stops growing. As an important information repository, email is expected to be much more than a simple communication channel. In particular, people are using email to perform tasks management and collaboration management. However, since email has large volume and is designed as a communications application, it's born with characteristics of lightweight and casualness. Using current email management tools, people have difficulties in organizing and managing their email data. Most importantly they have problems in using email to perform tasks management. In this paper, we propose a framework of entity-based task extraction from emails - TEXEM, which transfers plain-text email messages by entity-based clustering into collection of events and tasks and evaluates their importance after that. TEXEM can assist users to process their emails and thus improve working efficiency.

Keywords task extraction; entity identification; email management; clustering

1. 概述

近年来, 个人数据空间[1]的研究得到了越来越多的关注, 而电子邮件作为个人数据的重要来源, 包含着丰富的信息, 在个人数据空间的研究中占据着重要的地位。电子邮件现在被用来进行包括存档管理、任务管理以及联系人管理在内的各项个人信息管理任务[2]——Email 已经严重超载[3]。当今的

邮件管理工具, 如 Outlook[4]等, 对这些扩展需求的支持还远远不能令人们满意。

任务管理已经成为电子邮件的一个重要功能。在商业和个人来往的邮件中包含着大量的任务和事件信息, 完全手工地依靠阅读纯文本内容去处理这种类型的邮件通常会占用人们可观的时间和精力。

因此, 自动的邮件任务提取变得极为迫切。

我们观察到电子邮件具有很强的结构化信息, 如发件人和收件人等, 目前的方法[5, 6, 7]忽略了这种结构信息, 只是基于自然语言处理对邮件纯文本来提取任务, 存在一定的局限性。本文提出一种基于实体的邮件任务提取框架——TEXEM (Task EXtraction from EMails), 利用电子邮件的结构化信息, 结合邮件正文中重要的实体信息, 采用一种基于事件聚类的方法提取任务, 并且对用户的任务采用一种基于向量空间模型的算法进行重要性评估。

该工作的贡献包括三方面。首先, 我们提出了一种邮件任务提取框架; 其次, 我们给出了一种结合邮件结构和实体聚类的任务构造方法; 最后, 我们使用任务向量空间算法对任务重要性进行评估。

本文的组织结构如下: 第二部分介绍本文的相关工作; 第三部分介绍 TEXEM 的框架结构; 接下来对 TEXEM 处理框架中的核心问题——事件提取, 任务构造和任务重要性评估进行详细阐述; 然后是实验评估; 最后进行总结以及讨论未来的工作。

2. 相关工作

电子邮件中的任务提取和管理受到了研究界和工业界的高度关注, 出现了很多相关的工作, 我们在此列举一些具有代表性的工作。

在商用领域, 有一些邮件服务提供商, 如 Gmail[8]和 Windows Live Mail[9], 可以将邮件中具有较规范形式的事件信息提取出来, 例如图 1 所示提取出的电子邮件片段, 其中关于某次会议的时间和内容是具有较为规范的结构化的。但是经过观察我们发现, 电子邮件中大部分的事件信息都不具有规范的形式, 而是通过普通的语句来表达的, 如: “I will meet you at my office next Tuesday”。这种形式的事件和任务信息在邮件中占到了很大的比例, 但是目前的商用产品尚不能处理这样的情况。

在研究方面, 也有一些关于邮件任务管理和任务抽取的工作[5, 6, 10, 11, 12]。Whittaker 等人在[11]讨论了邮件中的任务管理在当前形势下的重要性以及一些解决思路。Bellotti 等人在[6]中探讨了一种以任务管理为中心的邮件管理工具。Gwizdka 等人在[7]中提出了一种以任务为中心展现电子邮件收件箱的方法。Almgren 和 Berglund 在[5]中提出了一种从邮件中抽取学术会议通知的方法, 该方法先将含有学术会议信息的邮件从其他邮件中隔离出来, 然后利用一些信息抽取技术抽取时间、地点以及主讲人等关键信息。Dalli 在[12]中提出了一种将邮件处理集成到个人信息管理框架中的方法, 目的是给用

户提供一种无缝的管理体验。Stern 在[13]中提出了一种从邮件中提取和解析时间信息的方法。

- Project Progress Overview by Jidong Chen 10:00 - 10:10.
- Technical Report by Qiong Wu 10:10 - 10:40.
- Experimental Analysis by Xiangyu Zhang 10:40 - 10:50.
- Technical Report and Demo on by Yukun Li 10:50 - 11:20
- Demo by Yubo Kou 11:20 - 11:30.
- Discussions for next steps 11:30 - 12:00.

图 1 一个具有事件信息的邮件片段

现有的事件和任务提取的方法都只考虑了邮件内容本身, 没有充分利用邮件中的结构信息。本文采用一种基于实体的聚类方法进行任务的抽取, 该方法将邮件中重要的结构信息和实体信息加以考虑, 能够更有效地提取出邮件中的任务信息, 并对其重要性评估。

3. TEXEM 系统框架

TEXEM 的系统框架如图 2 所示, 可以分为三个部分: 事件提取, 任务构造和任务重要性估计。

首先将具有同一主题线索的邮件进行聚类, 之后对每个线索聚类后的邮件进行词性标注和实体标注。我们利用标注得到的实体进行事件提取和构造。

从邮件对话中提取出事件后, 需要根据这些事件来构造用户的任务, 这里包括两个步骤: 实体识别和基于实体的事件聚类。

构造出邮件中每个所有者实体的任务后, 我们对任务中的事件进行重要性评估和排序。

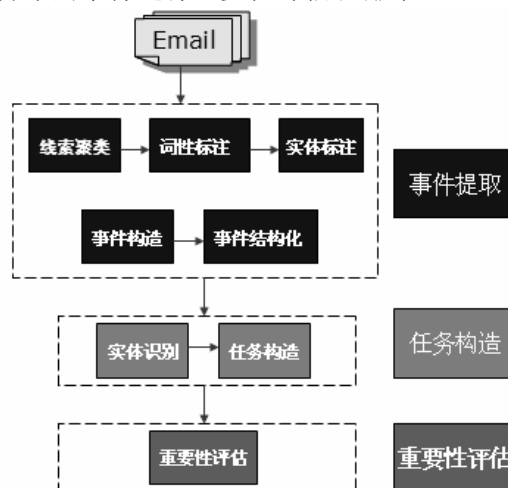


图 2 TEXEM 系统框架

4. 事件提取

4.1 事件提取

事件提取的处理流程如图 3 所示, 我们首先根据邮件的主题对邮箱中的邮件进行聚类, 形成多个

邮件线索，又称为邮件对话。之后以邮件对话为单位，对每个对话进行词性标注和实体标注（标识实意动词，标识人名、地名、时间等实体）。这两步中我们分别借助了 PoSTagger[14]和 ANNIE[15]这两个工具，经过这两步处理之后我们得到了构造事件需要的实体——邮件中的部分动词和名词，而且名词中作为时间、地点和人物的实体都被分别标识了出来，作为构造事件的原材料。

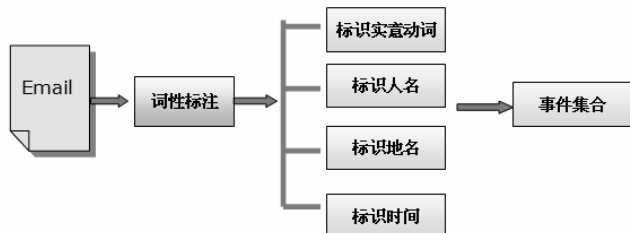


图 3 TEXEM 中事件提取流程

当得到构造事件所需的实体信息后，我们按照如下定义和结构来构造事件：

定义一（事件）：一个句子或者一个句子的一个分句包含一个事件，一个事件具有如下结构：

Event = {O, V, T, L, P, D}，其中：

- O：事件的所有者。
- V：事件的动词集合。
- T：事件发生的时间。
- L：事件发生的地点。
- P：事件涉及的人实体的集合。
- D：事件中的其他关键词。

基于以上的事件结构，我们将标注得到的实体按照词性填入到事件的相应部分。其中事件所有者实体 O 的处理会在第 5 部分任务构造时详细介绍。如下是一个事件的例子：

原句：I will invite Tom and Jerry to dinner today.
构造的事件：

E = {I, invite, today, NULL, [Tom, Jerry], dinner}

在某些情况下，一个句子中可能会缺失某些元素，这些情况该如何处理呢？

本文的处理原则是：如果句子中缺失 T、L、P 或 D 中的一个，那么就将这些元素置为 NULL；如果缺失的是 O，我们会对事件的所有者进行推测，具体方法会在第 5 部分阐述；如果缺失的是 V，那么本文认为这句话不含有事件，但是句子中的其他信息仍然是有用的，需要对事件进行合并。

4.2 事件合并和分解

如果一个句子中不含有实意动词（be 动词不是实意动词），则认为这句话中没有事件，需要将这个句子中的信息与离他最近的事件进行合并，并把这

个句子中其他成分合并后到对应部分中去。

另外，当句子中含有多个分句时，这个句子中的事件可能需要进行分解，分解的结果是每个事件对应句子中的一个分句。

5. 基于事件聚类的任务构造

经过观察，我们注意到电子邮件中实体的概念十分明显，一封邮件的内容通常都是围绕几个实体——尤其是人实体——展开的，如发件人和收件人就是很重要的实体。因此，有必要对上一步提取出来的事件以实体为单位进行聚类。

定义二（任务）：一封邮件的全部事件中所有者实体相同的事件集合构成这个所有者的任务。

从任务的定义可以看出，构造任务的关键就是要识别出每个事件的所有者，再对具有相同所有者实体的事件进行聚类。

5.1 实体识别

实体识别就是识别每一个事件的所有者，其具体过程是在处理每一个句子的时候标识出这个句子的主语——具体的人名、组织名或者一个代词，这个过程其实是在上面的事件构造时同时完成的。但是可能会出现句子中没有主语的祈使句，比如：

Send me those papers about email processing.
这句话中没有主语，所以我们推测其所有者为“你”，这是因为省略主语的情况大多数都是祈使句。

5.2 基于实体的事件聚类

识别出每一个事件的所有者之后，将这些事件按照它们所属的实体进行聚类来得到实体的任务。

图 4 所示的算法通过对事件所有者的判断，将属于实体“我”和“你”的事件分别进行了聚类，将其余事件暂时分组到一起。

```

(1) myEvents=empty;
(2) yourEvents= empty;
(3) otherEvents=empty;
(4) For each event e of the email thread
(5) If (owner of e is 'I' or 'We')
(6)   add e to myEvents;
(7) else if (owner of e is 'You')
(8)   add e to yourEvents;
(9) else
(10)  add e to otherEvents;
    
```

图 4 事件聚类的第一步

图 5 的算法对属于其他实体的事件依次进行判

断, 如果该事件的主语是一个人名, 那么就为其建立一个新的事件集合, 并将该事件加入; 如果该事件的主语不是一个人名, 而是人称代词, 我们将其加入到离其最近的实体的事件集合中去。最后, 将属于同一个人的多个事件集合并, 得到最终结果。

经过上面算法的处理, 属于其他人的事件被进一步细分, 我们得到了属于每个实体的事件集合, 也就是该实体的任务。

```

(1)  thirdPerson = unknown;
(2)  currentEventSet= new event set;
(3)  currentEventSet.owner = thirdPerson;
(4)  For each event e in otherEvents, from the fist one to the last one;
(5)  If owner of e is a person name, namely n
(6)  { thirdPerson = n;
(7)    currentEventSet = new event set;
(8)    currentEventSet.owner=thirdPerson;
(9)    add e to currentEventSet; }
(10) else {
(11)   add e to currentEventSet; }
(12) merge event sets with common owner;
    
```

图 5 事件聚类的第二步

6. 任务空间向量

任务构造过程提取出了邮件中每个人的任务, 每个人的任务由属于他/她的事件集合构成。我们认为一个任务中的事件是具有不同的重要性的, 所以我们构造任务空间向量对任务中的事件进行重要性估计, 并按照重要性对任务中的事件进行排序。

我们进行重要性估计的任务向量空间算法以向量空间模型中 TF-IDF 为基础。该算法以 TF-IDF 为基础对事件中词的权重进行了调整。

我们定义一个事件 e 的重要性得分如下:

$$\text{Score}(e) = \text{Mean}(\text{weight}(\text{term}))$$

其中 Mean 为一个均值计算函数, weight 为一个词的权值, term 为该事件中除所有者之外的其他部分中所有的词——动词以及其他名词。词的权重的计算我们采用以下的公式:

$$\text{Score}(\text{term}) = \text{factor} * (\text{TF}(\text{term}) * \log(\text{N}/\text{DF}(\text{term})))$$

其中 TF(term)和 DF(term)分别代表这个 term 在这个邮件对话中出现的次数和包含这个 term 的邮件对话的个数, N 为所有邮件对话的总数。factor 是一个不小于 1 的系数, 如果 term 是出现在邮件题目或者收件人、发件人等结构信息中的话, factor 取大于 1 的数, 在我们的实验中我们取为 2; 如果 term 不出现在任何结构信息中, factor 取 1。

按照以上方法对事件的重要性得分进行计算后, 我们对任务中的事件进行排序。

7. 实验评估

我们在实验中使用 Enron 邮件数据集[16], 其总大小大约为 1.2GB, 包含 150 个用户的共计 500 000 封邮件, 是目前最为通用的邮件数据集。实验的处理过程如图 6 所示。

图 7 所示的是一封示例邮件及处理后的结果, 左边是一封原始邮件, 右边是对其用我们的方法进行处理后的结果。从图中我们可以看出, 邮件中提到的三个事件被有效的提取了出来, 分别被构造成了属于“你”和“我”的任务, 而且对任务中的事件进行了重要性估计。

但是从图 7 的例子中也可以发现, 邮件中存在很多语言不规范的情况, 如用 i 代替 I, 用 u 代替 you, 用 pls 代替 please 等等, 对于这些情况 PoStagger 和 ANNIE 是无法处理的。而且邮件中还会出现广告等噪音数据, 如图 7 中原始邮件的最下方内容, 包括广告在内的噪音数据的存在对任务提取的质量会产生影响。

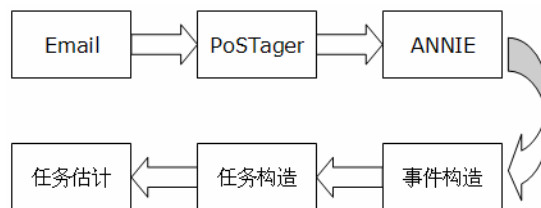


图 6 实验处理流程

```

Task for I:
NO.1 Importance Score: 4.0771135605671835
Action(s) : [scheduled, come, regards]
Element(s) : [gautan]
Time : null
Location : US
People Involved : []

dear bhaiya

i was making a list of all my contact
addresses in USA.could u pls send ur
complete address& ph/mobile.

NO.2 Importance Score: 3.18518880132031
Action(s) : [making]
Element(s) : [list, contact, addresses]
Time : null
Location : USA
People Involved : []

i am scheduled to come to US on 25th.
regards
gautan

Task for You:
NO.3 Importance Score: 3.5608222634318105
Action(s) : [send]
Element(s) : [address, ph\]
Time : null
Location : null

Get your FREE download of MSN Explore
r at http://explorer.msn.com/intl.asp
    
```

图 7 邮件任务提取实例

所以, 针对上述这些噪音以及其他特殊的情况, 我们制定了一系列的规则对邮件进行预处理, 以加强抽取任务的质量, 主要的规则包括:

1. 对于电子邮件中常见的缩写进行预处理, 转换为规则的形式, 如 i, u, pls 等。
2. 根据邮件中广告信息的出现规律, 比如常常以横线分割, 并且出现在邮件的末尾, 来制定规则对邮件中的噪音信息进行过滤。
3. 对于 let 这个特殊的实意动词的处理。由于 let

这个词一般在其后会有另一个实意动词，而由于 let 的含义的特殊性，所以我们认为 let 后面跟着的实体才是该事件的所有者，而不是 let 之前的主语，类似的，let 后面的那个动词才是属于该事件的动作元素的，而不是 let。

4. 有一些邮件，如新闻邮件等，其目的不是任务或事件交流，这一类邮件的特点一般是长度较长，所以我们在预处理时也将这一类邮件过滤。

经过实验证实，这些规则的引入对提高任务的质量具有很好的效果。

为了评估我们方法的有效性，本文对数据集中某个用户的 600 封邮件进行了任务提取实验，并对结果进行了分析。由于研究界对于事件和任务提取目前还没有统一的评估基准，所以本文以用户人工评估结果为基准进行测试。

实验评估采取的方法是先对上述的数据集中的用户的邮件数据进行任务提取处理，然后请 10 个人来模拟用户对这些邮件进行手工的任务提取和评估，处理的方式是从邮件中为每一句话提取事件，并构造每个人的任务，然后按照其重要度对任务中的事件进行排序。最后将这两者的处理结果进行比较，我们通过以下三个方面的度量进行比较和评估。

任务提取的召回率：手工处理提取出的事件数量为 a，程序处理得出的结果为 b，则召回率为 b/a。

任务重要性评估的准确率：手工处理对某个任务中事件的排序为 $Q=e_1e_2 \dots e_n$ ，程序处理的结果为 $Q'=e_1'e_2' \dots e_m'$ ，n 和 m 可能不同，则准确率为 $Distance(Q, Q')/(2n)$ ，其中 $Distance(Q, Q')$ 为序列 Q 和序列 Q' 的编辑距离。

实体识别的准确率：在程序提取出的事件中，能够被正确识别出事件所有者的事件数量与事件总量的比值。正确的标准为手工处理的结果。

我们对处理结果取平均数，结果如表 1 所示：

任务提取召回率	73.27%
任务评估准确率	54.49%
实体识别准确率	66.48%

表 1 TEXEM 实验结果

从表 1 的数据中可以看出，由于我们将每个句子都作为一个处理单位进行处理，所以任务提取的召回率是比较高的，但是由于电子邮件的随意性，以及我们使用的自然语言处理工具的局限，仍然会有一部分事件信息无法有效提取出来。

任务评估的准确率相比召回率要低一些。我们在评估时利用了邮件中的结构信息，如发件人、收件人和题目，但是由于自然语言处理方面的局限性，

准确率会受到影响，综合来看结果还是可以接受的。

实体识别的准确率也是不错的，但是由于电子邮件中语言使用的随意性及其多样化，可以看到距离使用户满意还有相当的距离。

8. 结束语

本文提出了一种基于实体的邮件任务提取框架 TEXEM，使用基于实体聚类的方法将邮件转化为任务的集合，并且对任务进行了重要性评估。TEXEM 提取出的任务结果呈现给了用户一个全新的邮件管理角度，可以对邮件管理起到很好辅助作用。

在以后的工作中，我们希望能够利用更多邮件中的元信息。我们还将考虑对事件进行相似度的计算，通过相似事件的合并来提高反复出现的事件的重要度。另外我们还需要考虑如何将处理结果作为辅助工具更好的呈现给用户。

参考文献

- [1] Michael J. Franklin, Alon Y. Halevy, David Maier: From databases to dataspace: a new abstraction for information management. SIGMOD Record 34(4) 2005: 27-33
- [2] Whittaker, S. Bellotti, V., and Gwizdka, J. Email in Personal Information Management. In Communications of the ACM 49(1): 68-73, 2006.
- [3] S. Whittaker, C. L. Sidner: Email Overload: Exploring Personal Information Management of Email. CHI 1996: 276-283
- [4] Outlook www.microsoft.com/outlook/
- [5] Almgren, Magnus and J. Berglund. Information Extraction of Seminar Information. 2000.<http://nlp.stanford.edu/courses/cs224n/2000/berglund/report.pdf>
- [6] Bellotti, V., Duchenaus, N., Howard, M., and Smith, I. Taking email to task: the design and evaluation of a task management centered email tool. In Proceedings Of Computer Human Interaction, 2003: 345-352
- [7] Gwizdka, J. Reinventing the Inbox. Supporting Task Management of Pending Tasks in Email. In Proceedings of Conference on Human Computer Interaction, 2002: 550-551
- [8] Gmail <http://gmail.com/>
- [9] Windows Live Mail <http://mail.live.com/>
- [10] J. A. Black, N. Ranjan Automated Event Extraction from Email <http://nlp.stanford.edu/courses/cs224n/2004/jblack-final-report.pdf>
- [11] Whittaker, S.. Supporting Collaborative Task Management in Email. Human-Computer Interaction, Volume 20, Issue 1 & 2 June 2005 , pages 49 - 88.
- [12] Dalli, Angelo. Automated Email Integration with Personal Information Management Applications. 2004. <http://www.cs.bham.ac.uk/~mgl/cluk/papers/dalli.pdf>

- [13] Mia K. Stern. Identifying and Understanding Dates and Times in Email IBM Technical Report RC-22875, 2003
- [14] PoSTagger <http://nlp.stanford.edu/software/tagger.shtml>
- [15] ANNIE <http://gate.ac.uk/ie/annie.html>
- [16] Enron Email Dataset. <http://www.cs.cmu.edu/~enron/>

张相於, 男, 1986 年生, 硕士研究生, 研究方向: 数据空间, 个人数据管理。

陈继东, 男, 1978 年生, 博士, 现为 EMC 中国实验室研究员, 研究方向: 个人数据管理, 知识管理, 知识智能。

李玉坤, 男, 1969 年生, 博士研究生, 研究方向: 数据空间, 个人信息管理。

孟小峰, 男, 1964 年生, 教授, 研究领域: Web 数据集成, XML 数据库, 移动数据管理。

论文负责人联系方式

学校: 中国人民大学

姓名: 张相於

电话: 62512334

手机: 13811064711

邮寄地址: 中国人民大学 红一楼 243 室

邮编: 100872

Email: zhangxy@live.com