一种基于邮件列表的软件问答信息抽取方法

罗宇翔 邹艳珍 金 庸 谢 冰
(北京大学信息科学技术学院软件所 北京 1000871) (高可信软件技术教育部重点实验室 北京 100871)

摘 要 开源项目通常会提供邮件列表来帮助用户更好地理解和使用开源项目。但由于邮件的数量巨大、邮件内容组织繁杂、问题不明确、答案定位困难等问题,用户在邮件查询过程中定位一个特定的软件问答信息要花费大量的时间和精力。为此,提出一种基于邮件列表的软件问答信息抽取方法。该方法通过对邮件的简单分类与标注,实现自动的问题句抽取和答案邮件选取,从而提升了用户进行邮件列表查询以及开源软件项目学习的效率。最后,通过实验验证了该方法的有效性。

关键词 软件复用,数据挖掘,邮件列表,软件问答

中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.12.005

Mailing List Based QA Information Extraction Approach

LUO Yu-xiang ZOU Yan-zhen JIN Yong XIE Bing
(School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China)
(Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing 100871, China)

Abstract Open source projects often provide mailing lists to help users better understand and use open source software. However, developers often spend a lot of time to retrieve the emails when they want to find a special answer, because there are a huge number of emails with unclear question and complex organization. User usually take a lot of email conversations before they get a right answer. In the paper, we proposed and implemented a question & answer information extraction approach based on open source software's mailing list. It can automatically extract the question sentence and the corresponding best answer from the emails, which can help users search mailing list and learn open source software more effectively. We also did some experiments to verify the availability and the efficiency of our approach.

Keywords Software reuse, Data mining, Mailing list, Software question & answer

近年来,随着软件复用技术的逐渐成熟以及 Internet 上开源软件的逐渐增多,软件开发人员在项目开发的过程中越来越多地复用开源软件[1]。同时,一些知名的开源社区如 Apache、Eclipse Foundation 等会在其开源项目的官网上提供诸如用户手册、FAQ 页面、邮件列表等信息来帮助用户更好地理解和使用开源软件。

邮件列表中蕴含着丰富的软件问答信息,有利于软件复用者理解软件资源的功能和使用方法[1]。然而,通常开源软件项目的邮件数量巨大,邮件内容十分繁杂,使得用户在邮件列表中查询特定问题变得十分困难。本文对开源社区 Apache Software Foundation 上两个知名开源项目 Lucene 和 Tomcat 的邮件数量作了一个简单统计:从 2001 年 9 月到 2013 年 11 月,Lucene 项目的邮件数量是 48626 个;而 Tomcat 项目的邮件数量达到了 134224 个,约占 1290M。同时,通常邮件内容是纯文本格式的,少数包含图片,要从数百兆甚至千兆的文本中查询特定的问题和答案是非常困难的。

为此,本文拟提出一种基于邮件列表的软件问答信息抽取方法。其主要面对以下 3 点技术挑战。

首先,如何准确识别和分类邮件内容?根据邮件在会话中的位置,可以将邮件分为提问邮件、答复邮件和回应邮件。而邮件内容的自由度非常大,经常包含一些与问答无关的其他类型信息,如邮件开头的问候语以及邮件结尾的引用和签名。邮件中的问题常常附带相关代码或异常文本片段。要想从内容繁杂的邮件内容中抽取出有用的问答信息,需要首先利用自然语言处理的技术对邮件文本段落进行良好的分类。

其次,如何从邮件会话中抽取问题?由于邮件文本的非正式性,邮件中的提问并不一定以疑问标点结尾,并且问题可能以陈述句式提出,使用传统的基于疑问词 5W1H 以及疑问标点的方法来抽取问题句的效果不佳,因此需要一种合适的方法可以依据邮件中提问的特征来抽取问题句。

再次,如何在众多的答复邮件中定位问题的最佳答案邮件?在一个邮件会话中通常会包含多个对提问邮件的回复。对于用户 A 的提问,可能 B 引用了 A 的问题,并做出回答。然而 A 对 B 的答案不是十分满意,最终 C 给出了 A 满意的答

到稿日期:2015-02-27 返修日期:2015-04-19 本文受国家高技术研究发展计划(863)(2013AA01A605),国家重点基础研究发展规划(973)(2011CB302604),国家自然科学基金(61103024)资助。

罗宇翔(1991-),男,硕士生,主要研究领域为软件工程、软件复用技术等,E-mail:zouyz@sei.pku.edu.cn;邹艳珍(1976-),女,副教授,主要研究领域为软件工程、软件数据挖掘、自然语言检索技术等(通信作者);金 庸(1990-),男,硕士生,主要研究领域为软件工程、软件复用技术等;谢 冰(1970-),男,教授,博士生导师,主要研究领域为软件工程、人工智能、形式化方法等。

案。因此,对提问邮件中问题的解答需要从与提问邮件相关的众多答复邮件中提取出来。但由于邮件内容中夹杂着冗余的邮件引用等噪音,使得仅仅依靠传统的文本检索的方法很难有效定位到满意答案。如何通过邮件会话的一些结构、内容特性来帮助定位答案信息是本文需要解决的关键问题之一。

1 问答信息抽取方法

基于上述分析,本文提出并实现了一种基于邮件列表的软件问答信息抽取方法。如图 1 所示,该方法主要包括以下 4 个部分。

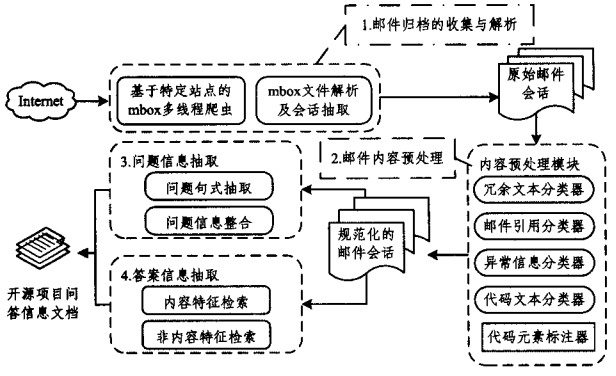

图 1 基于邮件列表的软件问答信息抽取方法

1)从指定站点收集和整理开源项目邮件。从邮件存储服务器爬取开源项目邮件归档文件,将邮件归档文件中的邮件按照讨论的主题组织成邮件会话的形式。(注:这部分工作较为简单,本文重点介绍后几个部分。)

2)解析邮件内容,对邮件内容进行预处理——将邮件内容分段,按照内容特征将段落分为冗余文本、邮件引用、代码文本、异常信息和正文信息等 5 类,并标注出邮件正文信息中的代码元素。

3)抽取提问邮件中的问题信息。抽取提问邮件中的问题句,并将问题句及与其相关的代码文本段落和异常信息段落整合成问题信息。

4)抽取邮件会话中的答案信息。基于内容特征从邮件会话中选取候选答案邮件,然后基于非内容特征核算权值,选取最佳答案邮件作为邮件会话的答案信息。

1.1 邮件分类与标注

邮件内容预处理包括邮件内容分类以及代码元素标注两部分。本文通过调研将开源项目的邮件正文内容大致可以分为以下 5 类文本:1)冗余文本段落,包括邮件中的问候语、感谢语、邮件签名等无用信息;2)邮件引用段落,在邮件中经常会出现引用之前邮件文本的情况;3)异常信息段落,开发人员在提问时通常会将出现问题的异常信息粘贴,异常信息可以帮助专业人士更高效地定位和解决问题;4)代码文本段落,开发人员在提问或者回答时会贴上相关的代码片段,可能是一段正确代码示例,也可能是一段有 bug 的代码;5)正文信息段落,除去以上 4 种类型之外的内容即为邮件的正文信息段落。前 4 类文本在邮件中有一定的规律可循,本文以这些规律作为特征构建分类器,以段落为基本单位对邮件内容进行分类。由于分类的次序很重要,本文采用了依次识别噪音文本、邮件引用、异常文本信息和代码文本的策略,从而可以减小噪音文本对后面内容识别的干扰,也减少了将部分文本识别为代码的误差。

由于代码元素在文本中有着比其他词汇更丰富的含义,标注出开源项目邮件正文信息中的代码元素对问答信息的抽取有着很大帮助。本文使用 Eclipse JDT 提供的 ASTParser 对项目 java 源代码进行了解析,并利用解析后的结果和camel-Case命名规则分别实现了对正文信息中代码元素的标注。

1.2 问题句的抽取

问题句通常指向整个邮件会话所讨论的核心。本文采用数据挖掘领域的序列模式挖掘算法来挖掘问题句式的模式,然后将提问邮件中的句子与问题句式进行匹配,以此选取问题句。

本文首先分别从 Lucene 和 Tomcat 两个项目的邮件列表中随机选取 1000 个邮件会话,由 7 名计算系学生(3 名本科生,4 名硕士生)从邮件会话中标注出问题句。问题句是提问邮件中表达作者疑问或者需求的一句话,问题句可以是疑问句也可以是陈述句。本文最终选取出 1973 个问题句。由于句子的词汇量巨大,如果以单词作为序列中的项,则会使得项集过大,因此考虑对句子进行词性标注,以词性标记作为序列项。本文采用 Standford Tagger 作为标注器,标注集为 Penn Treebank。同时,考虑到开源项目邮件的提问中经常会出现标题中的词汇和代码元素词汇,本文在 Penn Treebank 的基础上新增了标题词和代码词这两种词性标记。其次,本文选取了 PrefixSpan 算法[2]作为原型。相较于其他算法,PrefixSpan算法在选取候选频繁序列模式时不必多次重复扫描数据库,效率更高[3]。本文设定 PrefixSpan 算法中频繁序列在集合中的最少出现次数即最小支持度 min_sup 为 15,选取所挖掘出的序列模式中前 30 个频繁序列作为问题句式的模式。

在获取问题句式的模式之后,本文选取候选问题句中得分最高的句子作为最终的问题句。候选问题句的评分规则如下:1)每匹配一个模式,加 1 分;2)以问号结尾,加 1 分;3)候选句中出现标题词,加 1 分。可以看到,评分规则突出了标题在邮件中的重要性,并且在匹配同样序列模式的情况下,问句的优先级高于陈述句。

单一的问题句通常很难表达提问者的所有信息,与用户提问相关的文本片段如代码示例、异常信息、问题句的上下文等也属于问题信息的一部分。为此,本文还选取了问题句前后文本、问题句临近的代码文本片段和问题句临近的异常信息段落 3 类信息作为问题句的上下文,与问题句联合起来共同构成问题信息。

1.3 答案邮件的选取

答案邮件选取的实质是一个检索的过程,即在邮件会话的众多邮件中定位到最满意的答案邮件。本文工作中,首先使用基于内容特征的方法查找候选答案邮件,再使用基于非内容特征核算候选答案邮件的权值,选取权值最高的邮件作为该会话的答案。

在提问邮件中,标题中的内容词、问题句中的内容词和代码元素、异常文本段落中的异常信息词汇是 3 类比较重要的词汇,本文首先从这 3 类词中选取/构造查询词。为了提高查询的准确度,有必要对查询长度进行限制,过多的查询词会分散查询的主题[4]。本文发现平均长度为 5 的查询可取得较好的检索结果。另外,查询词之间的重要程度也不尽相同。例如问题句中的代码元素的重要性就要高于其他文本段落中的代码元素,因此应该赋予更高的权重。基于查询词,本文使用

向量空间模型(Vector Space Model,VSM)[5]计算文本与构造出的查询的相似度。该相似度即作为答复邮件和提问的关联程度,选择相似度最高的3个答复邮件作为候选邮件。

答复邮件中可能存在着对前文邮件的引用。而邮件引用的存在会干扰VSM模型的检索结果。本文选取了以下3个非内容特征来对候选答案邮件的权值进行核算,以此来帮助定位答案。

1)作者权威度:本文将邮件作者分为3类。项目开发人员的权重分值加1;答复邮件(不包括提问邮件)数目超过100的邮件作者定义为资深用户,其邮件的权重分值加0.5;其余作者作为普通用户,其邮件权重分值为0。

2)邮件位置:在邮件会话中,通常在较为靠后位置的答复邮件的内容的重要性更高[6]。本文按照发送时间对邮件会话中的答复邮件进行排序,从后往前邮件的权重分值依次加{1,0.5,0}。

3)答复邮件态度:在邮件会话中,若回复邮件末尾出现一些语气词如"good","cool",则将答复邮件视为被肯定;若回复邮件末尾仍是问句,则答复邮件被视为否定;其余情况视为中立。对于提问邮件,作者对答复邮件做出了肯定的回复,答复邮件权重加0.5分;对于提问邮件,作者对答复邮件提出了质疑,答复邮件权重减0.5分。

本文将3个维度的权值相加作为邮件的非内容特征得分,并选择得分最高者作为邮件会话的答案邮件。

2 实验

2.1 问题句的抽取

为了验证本文提出的问题句抽取方法,设计了如下实验:首先,从Apache Lucene和Apache Tomcat的邮件列表中随机抽取300个邮件会话,通过人工标注得到每一个邮件会话的提问邮件的问题句。其次,使用两种方法从准备的300个邮件会话中抽取问题句:1)使用传统基于疑问词和疑问标点的问题句抽取方法[7]进行问题句的提取;2)使用本文提出的序列模式挖掘的问题句抽取方法进行问题句的提取。

实验结果如表1所列,可以看出:基于疑问词和疑问标点的方法虽然有着不错的准确率(78.64%),但召回率比较低;相比而言,本文提出的基于序列模式的问题句挖掘方法对问题句的抽取不但有着较高的准确率(88.17%),而且召回率也达到了82%。因此,本文方法在实验中得到的F值要比基于疑问词和疑问标点的问题句抽取方法提高了20个百分点,从64.03%提高到了84.97%。

表1 问题句抽取方法实验对比

方法	准确率P	召回率R	F值
序列模式挖掘方法	88.17%	82%	84.97%
疑问词+疑问标点	78.64%	54%	64.03%

在上述实验中,基于疑问词和疑问标点的问题句抽取方法效果较差的主要原因是对邮件中以陈述句式提出的问题不能较好地识别。本文方法对300个提问邮件中的246个都准确地抽取到了问题句,取得了较好的效果。问题句抽取错误或者未抽取到问题句的原因主要有两个:1)邮件正文中包含多个问题。本文选择候选邮件中得分最高的句子作为问题句。2)邮件的问题句不匹配PrefixSpan算法挖掘出来的问题句模式。

2.2 答案的抽取

本文的答案邮件抽取方法在传统文本检索(图2中的Base方法)的基础上,增添了基于邮件作者权威度(Authority)、邮件位置(Position)、回应邮件态度(Attitude)3个非内容特征的衡量。进行了如下实验。

数据准备:从Lucene和Tomcat的邮件列表中随机抽取300个邮件会话,其中每个邮件会话至少包含2个邮件,平均每个邮件会话包含5.12个邮件。人工标注出每一个邮件会话的答案邮件。

方法:实验一共有5组:一组对照组,4组实验组。对照组是使用传统文本检索方法的Base组;实验组使用文本检索方法选取出候选答案,然后分别基于Authority(Au)、Position(Po)、Attitude(At)和Au+Po+At的特征对候选答案核算权值,进而选择最佳答案。在图2中,4个实验组分别表示为Base+Authority、Base+Position、Base+Attitude、Base+Au+Po+At。每一组的任务相同,即从准备的300个邮件会话中选取最佳答案邮件。

结果:实验计算了每一组方法在相同实验数据上的准确率和召回率,如图2所示。可以看出,Base方法的准确率和召回率不高,主要原因是在邮件中存在着大量的引用文本,增大了邮件之间的文本相似度,使得VSM模型的检索效果不佳;3个非内容特征对答案邮件抽取的准确率和召回率都有提高,Attitude特征效果不明显,这主要是由答案邮件的回应邮件并不多所造成。Authority特征对抽取方法的准确率和召回率都有较大的提高,说明邮件作者权威度对邮件质量确实有着很大的影响,资深用户以及项目开发人员的答复邮件通常更有价值。Position特征对答案抽取的效果提升最多,这表明在邮件会话中越靠后的邮件是答案的可能性更高。本文方法结合3种特征,对答案邮件抽取的准确率接近80%,相对于Base方法有了很大的提升。

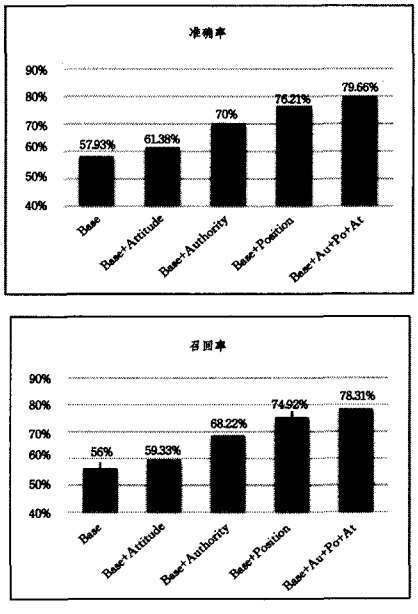

图2 答案邮件抽取方法实验对比

结束语 本文提出了一种基于邮件列表的软件问答信息抽取方法。基于对邮件内容的分类和代码元素的标注,提出了基于序列模式挖掘的问题句挖掘方法并采用内容特征和非内容特征结合的方法来定位答案邮件,从而实现邮件列表中

(下转第35页)

即使在滑动前后出现了属于相同图书的不同关键词，这也是对这本书另一个维度信息的表达，即实现了对图书多维信息的展示。

图 4 中，左图展示了应用为用户呈现的关键词。点击这些关键词可以选择查看相关关键词或查看相关图书。如果选择查看相关关键词，那么应用会为用户展示包含该关键词的图书对应的其他关键词；如果选择查看相关图书，那么会展示一个包含该关键词的图书列表。用户可以点击具体的图书来获得该图书的详细信息，见图 4 中右图。例如点击了图 4 左图中的“III-V semiconductor”关键词，选择展示相关图书，最终会呈现右图中有关这本书的详细信息。

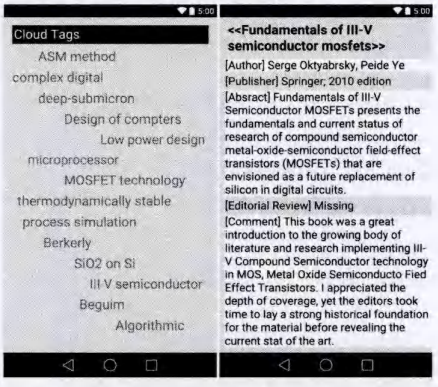

图 4　多维信息展示

5　案例评估

本文邀请了 21 名志愿者使用并评估该应用。

大多数志愿者认为该应用在判断用户面对书架中准确，但是仍有少数人认为不准确。经过仔细询问，发现这些人有时候会靠近一侧而远看另一侧的图书，此时对他们的定位就会出现偏差。

绝大部分的志愿者认为本应用为他们展示的关键词能帮助他们更好地理解该书架商品的信息。而个别的志愿者认为帮助不大，经了解，这些志愿者均为信息学院微电子系的学生，而实验区域的图书恰好包含了微电子领域的图书，因此他们认为展示的关键词没有提供额外信息。但是他们依然认为该应用能够帮助他们发现位于书架边缘容易忽视的图书；即使是非理工学院的参与者也能够根据关键词发现一些非纯技术的图书，例如《Politics and technology》。

结束语　本文设计并以案例形式研究、实现了一种面向浏览购物行为模式的室内 LBS 购书应用。该应用能够为在浏览模式下购物的用户提供图书的多维概览信息，帮助他们理解周围图书的内容，进而发现自己感兴趣的图书，产生购书的欲望。经过案例评估发现该应用能够增强用户的购物体验。

参 考 文 献

[1] Resatsch F, Karpischek S, Sandner U, et al. Mobile sales assistant: NFC for retailers[C]// Proceedings of the 9th International Conference on Human Computer Interaction with Mobile Devices and Services. ACM, 2007: 313-316

[2] Resatsch F, Sandner U, Leimeister J M, et al. Do Point of Sale RFID-Based Information Services Make a Difference? Analyzing Consumer Perceptions for Designing Smart Product Information Services in Retail Business[J]. Electronic Markets, 2008, 18(3): 216-231

[3] Von Reischach F, Guinard D, Michahelles F, et al. A mobile product recommendation system interacting with tagged products[C]// IEEE International Conference on Pervasive Computing and Communications, 2009 (PerCom 2009). IEEE, 2009: 1-6

[4] Black D, Clemmensen N J, Skov M B. Supporting the supermarket shopping experience through a context-aware shopping trolley[C]// Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open. ACM, 2009: 33-40

[5] Kallehave O, Skov M B, Tiainen N. Persuasion in situ: shopping for healthy food in supermarkets[C]// Proceedings of PINC 2011 Workshop at CHI. 2011

[6] Shekar S, Nair P, Helal A S. iGrocer: a ubiquitous and pervasive smart grocery shopping system[C]// Proceedings of the 2003 ACM Symposium on Applied Computing. ACM, 2003: 645-652

[7] Muñoz-Organero M, Muñoz-Merino P J, Delgado Kloos C. Using bluetooth to implement a pervasive indoor positioning system with minimal requirements at the application level[J]. Mobile Information Systems, 2012, 8(1): 73-82

[8] Witten I H, Paynter G W, Frank E, et al. KEA: Practical automatic keyphrase extraction[C]// Proceedings of the fourth ACM conference on Digital libraries. ACM, 1999: 254-255

（上接第 25 页）

问答信息的抽取。本文通过实验验证了方法的有效性。但是本文方法仅考虑了单一主题的邮件会话，并不适用于包含多个问题的多主题邮件会话。未来我们将针对这一不足进行深入研究。

参 考 文 献

[1] 金庸. 基于邮件列表的软件问答信息抽取工具的设计与实现[D]. 北京：北京大学，2014
Jin Yong. A design and implementation of software R&A extraction tool based on maillists[D]. Beijing: Peking University, 2014

[2] Fournier-Viger P. Spmf: A sequential pattern mining framework [OL]. http://www.philippe-fournier-viger.com/spmf, 2011

[3] 肖仁财. 序列模式挖掘算法研究与实现[D]. 南京：江苏大学，2007
Xiao Ren-cai. A research and implementation of sequential pattern mining algorithm[D]. Nanjing: Jiangsu University, 2007

[4] Belkin, Nicholas J, et al. Query length in interactive information retrieval[C]// Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2003

[5] Salton, Gerard, Wong A, et al. A vector space model for automatic indexing [J]. Communications of the ACM, 1975, 18(11): 613-620

[6] Cong G, Wang L, Lin C Y, et al. Finding question-answer pairs from online forums[C]// Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2008: 467-474

[7] Wang Kai, Chua T-S. Exploiting salient patterns for question detection and question retrieval in community-based question answering [C]// Proceedings of the 23rd International Conference on Computational Linguistics. 2010