软件外包项目工作日志质量评估方法①

姚承昊[1,2], 杜 晶[1], 肖俊超[1]

[1](中国科学院软件研究所 基础软件国家工程研究中心, 北京 100190)
[2](中国科学院大学, 北京 100190)

摘 要: 工作日志是软件外包项目监控项目进展的一个重要手段,它由工作人员填写汇报项目进展. 工作日志的质量一定程度上体现了过程执行的质量, 但是由于其数量庞大内容琐碎, 很难依靠人工完成查看和评估. 现有的研究对日志质量的关注较少, 因此本文提出一个自动化评估日志质量的方法, 该方法利用词法分析、依存句法分析、LDA 主题模型对历史日志数据进行分析和挖掘, 从结构、内容、主题相关性等方面选取质量特征, 通过专家小组标注的方法获得训练数据, 使用分类算法建立质量评估模型, 通过模型对日志质量进行自动化评估. 本文以一个国家核高基项目为案例, 实现了一个具有较高准确率的评估模型, 结果表明本文的方法能够合理的评估日志质量, 为外包单位提供有效的决策支持.

关键词: 软件外包监控; 工作日志; 文本挖掘; 质量评估

Method of Evaluating the Quality of Work Diary in Software Outsourcing Project

YAO Cheng-Hao[1,2], DU Jing[1], XIAO Jun-Chao[1]

[1](Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Science, Beijing 100190, China)

Abstract: Working diary is an important way of monitoring the progress of software outsourcing project. They are committed by project staff as a report of daily work. The quality of working diary can reflect procedure execution of project to some extent, but because the content of working diary is large and trivial, it is hard to check and evaluate rely only on manual work. Existing researches pay little attention to the quality of logs, so we proposed an automated method to evaluate the quality of working diary. Firstly, this method uses the lexical analysis, interdependence syntactic analysis and LDA model to mine and analysis the historical data. Secondly, we extract quality features from aspects such as structure, content, subject relevance, then collect training samples by artificial tagging. Lastly, we establish evaluation model using classification algorithm, through which we get the final evaluation of the work diary. This paper made a case study based on a national project and achieved a highly accurate evaluation model. The result shows that the method can effectively evaluate the quality of working diary, which helps to make decision for outsourcing department.

Key words: software outsourcing monitoring; work diary; text mining, quality evaluation

1 引言

软件外包在降低开发成本, 提高资源配置效率等方面具有很大的优势, 能使企业更加专注于核心能力和核心竞争力, 受到越来越多软件企业的青睐. 在外包中, 发包方不参与项目的开发, 无法直接获悉项目开发中产生的问题和偏差,具有较大的风险,因此对项目开发过程的监控显得尤为重要. 外包项目的监控重点是实际进度是否与计划相符, 承包方的投入是否充分, 工作成果质量是否合格等. 外包项目的监控不能像传统项目一样实时的监测到项目计划的参数如进度、工作量、费用、资源、工作成果等, 更多的是通过承包方提交的报告等来评估项目的情况, 例如人员

① 基金项目:核心电子器件、高端通用芯片及基础软件产品项目(2012ZX01039-004);国家自然科学基金(61432001,91218302)
收稿时间: 2015-02-05; 收到修改稿时间: 2015-04-02

工作日志, 课题进展周报等.

工作日志是一种细粒度的项目进展报告,由参与项目的工作人员提交, 供项目负责人、发包方管理者查看, 以日为频度总结和汇报当日工作情况. 项目进展报告描述项目团队在某一特定时间段工作完成情况, 主要形式为周进展报告和月进展报告, 在软件能力成熟度模型 CMMI[1], 精简并行过程 SPP[2]中对其格式和内容都有明确要求. 项目管理者通过周报、月报能够从整体了解项目的基本情况, 但还有潜在的问题有待挖掘. 研究者在[3]中提出了一种格式自由的进度报告形式, 由工作人员用简洁的语言实时报告项目的进度和遇到的问题, 工作日志与之相似, 是周报和月报的补充, 一方面提供更细粒度的信息, 另一方面由一线的工作人员填写反映的问题更真实.

日志的填写情况一定程度上体现了过程执行的质量, 日志内容丰富每天都有新的进展说明项目有活力, 产出也会有较高的保证, 相反日志内容简单重复项目就可能有潜在的风险. 通过对日志信息进行分析和评估能够发现项目实际进度是否产生偏差, 承包方的投入是否充分, 是否有潜在的风险等问题, 为外包监控和项目质量评估提供有效依据, 同时可以督促承包方对项目的投入. 日志报告包含大量的信息, 但相应的数量也十分庞大且内容琐碎, 难以通过人工的方式来对日志内容进行评估, 需要利用挖掘的方法来发现日志中影响质量的因素, 通过机器学习的方法学习影响因素和质量之间的关系, 根据关系建立评估模型, 进而自动化的评估日志质量.

为帮助外包单位更好地了解外包任务的实际进展, 进而评估项目的质量, 本文以外包软件项目工作日志为研究对象, 提出了一个分析和评估日志内容质量的方法. 该方法结合已有的文本挖掘和语言分析相关知识对工作日志质量的影响因素进行分析和挖掘, 首先将没有固定格式和内容要求的日志文本信息分解为容易分析的分词、词频等数据, 通过对历史日志数据的分析确定影响日志质量的特征. 然后, 根据项目需求对日志质量进行分级, 通过专家小组人工标记获得训练样本, 利用机器学习的分类算法在训练样本上学习质量特征和质量之间的关系, 选择最优的分类算法建立评估模型, 代替人工评估日志内容的质量. 通过一个国家核高基项目作为案例的实践, 建立了一个具有较高准确率的评估模型, 证明本文提出的方法能够合理的评估日志的质量.

2 相关研究

2.1 软件外包监控

软件外包(Software Outsourcing)是指发包方通过签订合同的形式, 将软件项目中的部分工作交给软件外包服务商(承包方)代工开发, 以获得高质量、低成本的软件产品的一种业务管理模式. 外包的实质是一种更加有效率的资源配置, 因此对外包的管理很重要, 外包的管理流程大概分为如下几步: 外包决策、选择承包商签订外包合同、监控外包开发过程、成果验收. 其中监控外包开发过程是外包管理中贯穿项目始终的主要管理内容, 是外包管理领域的一个研究重点.

成功的软件外包项目要求发包方必须跟踪和监控分包方的项目进展情况和关键的中间产品, 当承包方的进展和计划有较大偏离时做出适当的措施进行调整和纠错. 外包过程监控的主要手段是里程碑式评审和派遣外包经理. 里程碑式评审是在项目中设置若干个里程碑, 在里程碑处发包方对项目进行检查和评估. 外包经理是由发包方派遣长期进驻承包方单位, 专门负责监控和交流事宜[4].

对外包监控的研究主要集中在上述两种方法, 除此之外, 承包方定期提交的进展报告、技术报告等报告中也包含大量的有效信息,可以多角度深入的体现项目的实际情况, 对报告内容的挖掘和分析能够为外包监控提供很大的帮助, 但由于文本内容解析难度大原因缺乏有效的分析方法.

2.2 软件仓库文本挖掘

软件仓库挖掘(Mining Software Repositories)是对软件仓库中丰富内容进行分析和挖掘发现其中有价值信息的学科, 近年来得到充分的发展成为研究领域的一个热点. 源代码、版本档案等内容由于有固定格式易于解析得到广泛的研究, 而对于软件仓库中由自然语言组成的文本内容, 虽然也包含大量的有效信息, 但因其没结构随意不易于解析, 长期被研究者所忽略.

Alexander 和 Jane 等在[5]中阐述了对文本类型信息挖掘的意义, 同时证明了信息检索、机器学习等方法在文本产品信息挖掘中的有效性, 介绍了基本的文本挖掘方法. 此后, 缺陷描述、邮件、需求文档、产品说明书、设计文档等都得到较多的重视和研究, 在缺陷预测、需求跟踪、错误分析等领域取得一定的成绩[6–7].

但是研究的对象都是传统项目或者开源项目的文档, 对外包型项目中的文档关注较少.

2.3 自然语言处理相关知识

词法分析、句法分析是汉语分析技术的基础, 是本文对日志质量进行研究的基础. 词是最小的具有意义的独立语言成分, 但汉语以字为基本的书写单位, 词语之间没有明显的区分标记, 因此中文词法分析是中文信息处理的基础与关键. 词法分析的任务是将汉语句子划分成独立的词序列, 并对词进行词性标注, 目前词法分析效果较好较常用的是基于层叠隐马模型的汉语词法分析. 句法分析的输入是由词法分析输出的词序列, 输出是句子的句法结构, 是对词语语法功能的分析, 其面临的问题各种语言没有太大不同, 所采用的技术大体一致. 句法分析一般都依赖于某种语法体系, 不同的语法体系产生的句法结构形式不尽相同. 主要的句法结构形式有依存关系树, 句法树, 特征结构等. 目前应用最广泛的句法分析算法是 Tomita 算法[8]和 Chart 算法[9].

本文中利用 LDA 计算日志文本相似度, LDA 主题模型认为每个词都是属于某个主题的, 字面不同的词如果属于相同主题仍然有一定的相似性[12]. 基于 LDA 的文本相似度计算方法可以计算出文档的主题分布, 并用主题分布向量来计算文本的相似度, 将语义考虑在内, 相比于传统的只考虑字面的相似度计算方法更合理有效.

3 日志质量评估方法

3.1 方法简述

建立日志质量评估模型评估日志质量的流程如图 1 所示. 方法步骤如下:

① 数据处理: 首先将不符合基本要求的日志进行过滤. 然后构建分词工具、依存句法分析工具、LDA 文本相似度计算工具对日志进行处理和解析.

② 特征选取质量分级: 建立评估模型需要利用分类算法来学习人工评估的规则, 将质量影响因素转化为可以量化的特征, 通过特征来判断日志的质量, 模拟人工的评估. 因此, 需要对处理后的文本分析数据进行观察分析, 总结日志的特点和存在的问题, 根据分析确定影响日志质量的主要因素. 提取质量特征. 同时, 根据项目的需求对日志质量进行分级, 确定质量好坏的标准.

③ 抽样和人工标注: 建立分类评估模型需要有监督的学习, 本文采取人工评价的方法, 从日志数据中抽样出一部分数据, 根据之前确定的质量分级标准进行人工标注.

④ 特征验证: 利用标注数据集对潜在特征空间进行分析, 确定特征空间.

⑤ 训练模型和质量评估: 利用机器学习的方法以标注数据集作为训练样本, 学习人工标注的规律, 建立分类评估模型. 对于需要进行评估的日志, 通过数据处理的方法获得相应特征值, 以特征值为输入利用评估模型获得质量评分.

3.2 节到 3.6 节分别对上述步骤进行详细说明.

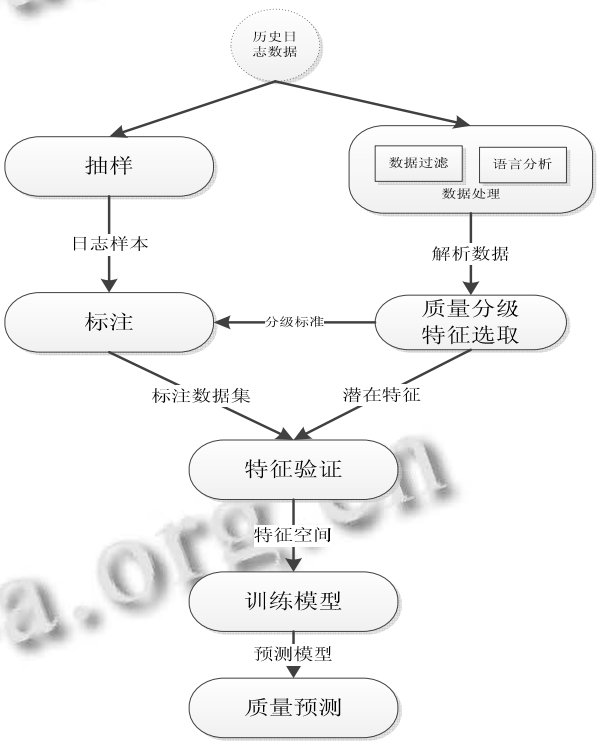

图 1 日志质量评估架构

3.2 数据处理

工作日志一般通过协同工作平台提交, 内容为工作人员对当日的工作内容或者遇到的问题的描述. 一般都比较紧简短, 例: “LSM 框架结构调研”, “查阅系统完整性度量机制原理的相关知识、掌握基本概念”. 数据处理包括两部分: 数据过滤和文本分析. 处理流程如图 2 所示.

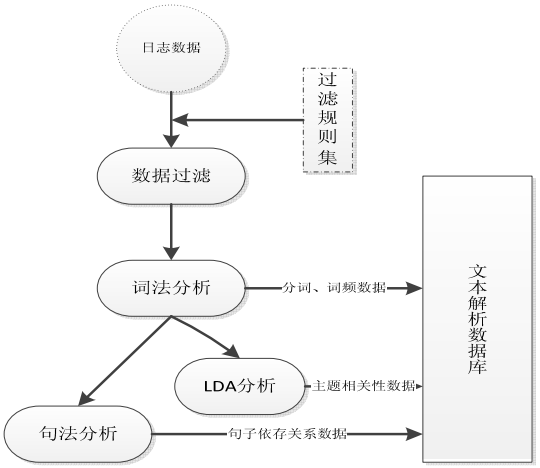

图 2 数据处理流程图

3.2.1 数据过滤

不同应用场景下对日志的要求不同, 但日志作为工作汇报有一些基本的要求必须满足, 不满足要求的日志直接确定为无效日志, 不参与质量评价, 通用的过滤规则有如下几种:

① 结构不完整: 日志作为工作情况的报告, 主语默认为报告人可以省略, 谓语和宾语对应必须要记录最基本的工作的行为和工作的对象信息, 不能省略. 因此, 缺少谓语或宾语的日志认为是无效的日志.

② 内容重复严重: 日志报告的一个重要特性是以日为频度, 如果连续多日的日志内容重复就失去了日志报告的意义. 考虑到工作人员遇到困难可能一天之内无法解决, 本文认为小于等于三天的日志是有效的, 大于三天的无效, 具体标准可根据应用场景适当调整.

③ 内容与任务无关: 因为日志数量比较庞大, 项目管理者难以一一查看, 有些工作人员会随意粘贴一些文本充当日志, 这部分日志也应该被过滤.

3.2.2 文本分析

日志是一种篇幅较短的文档, 内容由自然语言组成, 在挖掘和分析之前需要进行语言分析转化为易于解析和观察的数据形式. 文本分析主要包括词法分析, 句法分析, 文本相似度分析等通用分析过程, 数据过滤和文本分析后的数据以表的形式存储在数据库中, 供后续的使用. 文本分析基本步骤如下:

① 词法分析: 首先, 对日志文本进行分词处理, 使用中科院计算所汉语词法分析系统 ICTCLAS[11]提供的开源代码构建分词工具, 对日志文本进行分词和词性标注, 将句子转化为词和词性序列. 然后, 对分词结果进行统计, 获取需要的词数、词性数、词频等数据.

② 句法分析: 在词法分析的基础上, 使用哈尔滨工业大学语言技术平台[10]提供的依存句法分析服务构建句法分析工具, 句法分析以词法分析的结果为输入, 分析获取日志的词与词之间的依存关系. 通过依存关系找到日志的关键成分, 如中心谓语中心宾语等, 以帮助质量特征的获取.

③ LDA 分析: LDA 分析同样在词法分析的基础上进行, 利用 LDA 主题模型计算日志之间的相似度, 作为日志与主题相相关性的度量. 计算出日志的主题分布模型, 用主题的分布向量计算日志与日志之间的相似度, 然后进行加权平均等方法获得日志考虑语义的相似度数据.

3.3 特征选择提取和质量分级

3.3.1 特征选择

评估模型通过可量化的质量特征来判断日志质量高低, 质量特征越全面越能体现日志的特点, 建立的模型效果就越好, 因此日志质量特征的选取是建立评估模型的关键, 需要对经过处理的日志数据进行深入分析, 观察日志的特点, 总结日志存在的主要质量问题, 根据应用场景定义相关的特征. 特征主要从以下三个方面来考虑: 日志结构复杂性、内容准确性、主题相关性. 下面给出三者的详细解释和可能的影响特征.

① 结构复杂性: 日志结构复杂性是指日志句子结构的复杂性, 一般情况下, 结构复杂的日志对工作内容的描述更详细, 包含的有效信息更多, 质量也相应的高. 日志结构复杂性主要和日志中包含词数、词性数、各主要词性的词数等相关.

② 内容准确性: 结构的复杂性不能等价于日志的质量, 相同结构的日志不一定具有相同的质量, 例如“分析代码”和“分析 singal.h”, 结构虽然相同, 但后者的中心宾语“singal.h”定义更准确, 更能清晰的反映工作情况, 质量较高. 日志内容准确性主要和日志谓语中心语, 宾语中心语相关.

③ 主题相关性: 主题相关性是指日志和任务主题的相关程度, 如果内容和任务主题不相关, 结构复杂内容准确都没有意义. 日志主题相关性主要和考虑语义的日志文本相似度相关.

3.3.2 特征提取

选取特征的特征值需要对原始数据进一步加工获得, 这个过程就是特征提取的过程. 特征提取一般通过程序、统计工具等方法来完成, 不同的特征提取过程不尽相同, 需要根据具体情况设计相应的程序或者利用统计工具完成. 例如: 上节提到的与结构复杂性相关的日志各主要词性的词数. 可以通过程序读入分词后的日志数据, 对其进行一遍遍历, 记录下相应词性的标记出现的次数, 即可获悉此日志中各词性词的数量. 总词数、词性数等都可以通过此方法获得特征值. 内容准确性相关的中心词词频则需要对所有日志数据进行一遍遍历之后, 记下所有词的词频, 经过统计获得. 词频数值差距较大, 高的达到上万低的只有个位数, 需要进一步的离散化处理, 可以利用 SPSS 统计工具的相应方法完成.

3.3.3 质量分级

本课题研究的目的是对日志的质量进行评估, 但是日志本身是没有固有的质量评级的, 需要根据项目的需求和数据集的特征将日志进行质量分级, 分级的标准从上文中提到的日志、内容、主题三方面进行考虑. 按照日志的一般特性本文推荐将质量分为四个等级, 从低到高如表 1 所示. 在具体应用本文方法时, 可根据实际情况相应调整

表 1 日志质量分级

质量等级	标准
第一等级	拥有基本的谓语宾语结构简单的单句，中心语描述笼统，内容准确性差。
第二等级	拥有基本的谓宾结构，同时中心语准确性高，或者中心语准确性低但结构较复杂。
第三等级	结构为复杂的单句或者有递进、并列等关系的复句，且内容精确。
第四等级	在第三等级基础上，有多个句子，或者一个句子包含内容较多，内容丰富，撰写态度认真。

3.4 抽样和人工标注

利用分类算法训练分级评估模型, 需要有监督的进行学习, 对抽样的日志数据进行人工标注获得有质量评价的日志数据, 以之为训练数据集来训练评估模型.

① 抽样: 日志的抽样要均匀随机, 以保证抽样数据具仍然保持整体数据集的特征, 如果日志的抽取集中在某个时间段或者某几个人中, 日志数据呈现的特性会具有局限性, 建立的模型效果也会受到影响. 抽样从人员、时间、日志长度等维度进行综合考虑, 保证各类型各时间段的日志都有抽取. 样本的数量视项目具体情况而定, 样本越大建立的模型效果越好, 但人工标注工作量较大, 需要综合考虑效果和人力消耗.

② 标注: 标注以定义的日志质量分级为依据, 人工进行判断分级. 标注只看日志的文本内容, 隐藏其他统计数据, 避免被某些特征影响. 为保证标注数据的可用性, 要求多人对同一样本进行标注, 并对标注的结果进行 T 检验, 检验合格的数据才能用来训练模型.

3.5 质量特征验证

通过对日志数据分析选取的质量特征是潜在特征, 其有效性需要进一步的验证. 对潜在质量特征和日志质量分级结果进行相关性分析, 选择相关性显著的特征确定最终的特征空间.

3.6 训练模型和质量评估

日志的质量是离散的分级, 因此本文选择分类算法来建立评估模型. 通过分类算法学习质量特征和日志质量之间的关系和分辨质量高低的规则, 最后以公式、决策树等方式呈现. 不同的分类算法在不同的场景和数据集下表现不同, 本文的方法从决策树、朴素贝叶斯、贝叶斯网络、逻辑斯蒂、随机森林等分类算法中选择效果最优的算法来建立模型.

利用机器学习工具 Weka[13]训练样本, Weka 是一个专业的机器学习工具, 实现了的各种分类、聚类、回归等类型的机器学习算法. Weka 训练算法模型的过程简单易操作, 首先在工具中读入处理好的数据, 并选择数据中的目标特征, 即要预测的特征, 本文中就是专家标注的日志评分. 然后选择算法的种类(分类、回归、聚类), 在相应的种类下选择具体的算法, 例如: 选择分类算法中的决策树 J48, 如图 3 所示. 然后选择算法参数、验证方法(交叉验证)等等内容. 期间还可以先对数据进行过滤, 选择各个数据的数值类型(分类、连续), 进行一些统计等等, 不在这里一一介绍. 选择好所有的参数之后, 点击运行, 完成后即可获得算法的模型以及显示在界面右边的算法效果的详细分析.

本文通过模型的准确率, 相对误差绝对误差等指标来决定模型的优劣, 模型好坏的主要标准是准确率, 在准确率相似的情况下综合考虑其他几项指标. 建立模型的流程如图 4 所示.

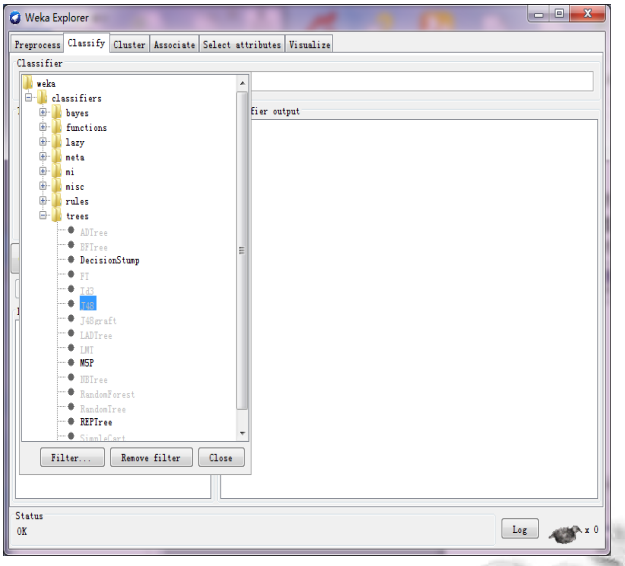

图 3　Weka 操作界面

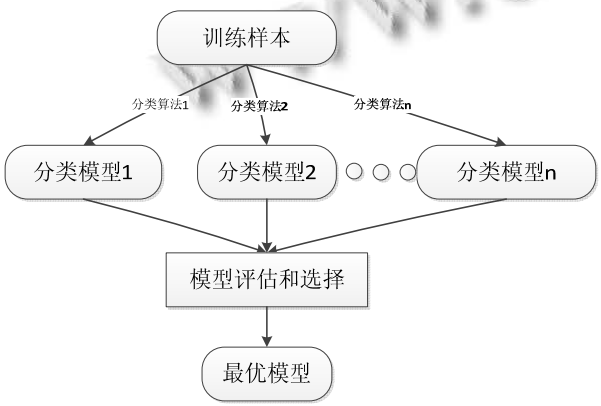

图 4　建模流程图

需要评估质量的日志通过训练集相同的特征提取过程获得特征值，将日志特征序列输入到模型中，模型通过一定的计算输出一个日志分类结果，这个分类结果就对应日志的质量等级，即可获得我们需要的日志评估结果.

4　案例研究

4.1　数据背景

本文的案例数据来源于一个外包型的国家核高基项目. 课题是一个综合性科研课题，主要是对 Android 和 Linux 操作系统源代码的分析和自主操作系统的研发，最终的产出有报告也有软件产品. 总体单位将课题分为若干个子课题，分给相关研究机构和组织来共同完成，通过一个协同工作平台监控和管理整个课题. 承包方需要定期在工作平台上提交报告和成果，工作平台上的报告是项目过程评估的主要依据. 人员工作日志是其中一个重要的报告，需要参与项目的工作人员在每个工作日提交，汇报当日工作内容.

在课题中对项目的评审分为两部分，一部分是季度评审中专家对项目的评分，一部分是根据项目提交的报告确定的过程评分，前者占最终评价的 70%，后者占 30%. 工作日志是过程评分中重要的一项，占过程评分的 1/3. 但是由于人力限制，项目对工作日志的评估只考虑是否提交，提交是否及时，没有考虑日志内容的质量，造成评估的不合理，有些单位虽然一直按时提交日志，但内容都很简单甚至一直重复，也得到了很高的过程评分. 为了更合理的评估工作日志，更好的评估项目的质量，我们用本文提出的方法对日志进行分析和评估，帮助项目管理组对项目的监控和项目质量的评估.

本文研究的对象是课题两年研发过程中工作平台上产生的所有日志数据，共有 21 家单位参与课题研发，子课题数 62 个，拆分任务数 8532 个，参与研发人员 815 人，日志报告数 95450 条.

4.2　日志数据处理

根据第三章中描述的方法对案例数据进行过滤，过滤掉基本信息不全和重复情况严重的日志共 29438 条. 信息不全的日志主要有“进行中”，“进展顺利”，“良好”，“XX%”等只笼统的记录任务进展状态，没有具体内容的日志，信息不全的日志大部分同时是重复的日志.

日志数据经过词法分析获得 801533 个词，各个词性的词频分布如图 5 所示. 通过文本分析我们发现日志数据的一些特点:

① 日志句子组成中，名词和动词的数量占较大比率，占总词数 29.8%和 39.2%.

② 日志篇幅较短，只有 2523 条日志包含多于 1 个句子，日志句子结构较为简单，包含词数小于等于 10 的日志占总日志数的 75%，词性数小于等于 5 个词性的日志占总数的 75.4%.

③词的词频和内容的准确性有一定的负相关关系，词频高的词都是概括性较高的词，如“代码”“系统”等，词频较低的特指性较强，例如“钩子函数”“MemoryComPaction”等.

④ 日志重复情况普遍，815 人中只有 78 人没有重复的日志，所有日志中重复出现的占 498.%，连续重复大于等于 5 天的日志有 28603 条.

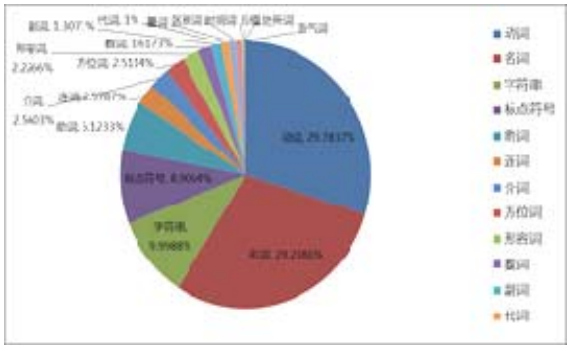

图 5 日志各词性词的数量比率

4.3 特征选取和质量分级

4.3.1 日志报告质量主要问题

由于缺乏对日志内容的监控, 案例中日志的内容存在较多的质量问题: 有很大一部分日志句子结构不完整, 不包含必须的谓语动词和宾语名词, 数量为 17744 条达到总日志数的 18.6%; 日志描述内容空泛, 有部分日志虽然描述了工作对象和工作行为, 但是用词宽泛, 包含信息量很少. 例如“分析代码”等; 日志内容重复情况严重.

4.3.2 日志质量特征选取

根据对日志的特点和主要质量问题的分析, 本文根据日志结构复杂性、内容准确性、主题相关性来选取特征, 具体如表 2 所示.

表 2 日志质量特征

维度	名称	描述
日志结构相关	Word_Num	词的总数
	WordType_Num	词性的总数
	Noun_Num	名词数
	Verb_Num	动词数
	Adjective_Num	形容词数
	String_Num	字符串数
	Auxiliary_Num	助词数
	Conjucnction_Num	连词数
	Prepsition_Num	介词数
日志内容准确性	Noun_frequency_Discretize	中心名词在所有日志中的词频
	Noun_frequency_user__Discretize	中心名词在个人日志中的词频
	Verb_frequency__Discretize	中心动词在所有日志中的词频
	Verb_frequency_user__Discretize	中心动词在个人日志中的词频
日志主题相关度	Similarity	日志和任务主题的相似度

4.3.3 日志质量分级

根据日志数据集的特征, 案例中将日志质量的等级按照第三章的标准分为四类, 四个等级中第二等级的日志数量最多, 形式如“阅读 Linux Kernel Development”, 是最常见的日志形式. 第一等级和第三等级日志数量相当, 前三等级的日志占日志数量的绝大部分, 第四等级日志较少, 是超出预期的高质量日志.

4.4 抽样和人工标注

根据案例数据的规模抽取了容量为 5000 的一个样本, 三个人进行同时进行标注. 对标注后的三个数据集两两进行配对样本 T 检验, 检验结果如表 3 所示, 两两之间的 P 值都远大于 0.01, 说明三次标注的结果没有明显的差异, 数据的标注结果是客观合理的. 标注后的日志评级分布整体符合正态分布, 如图 6 所示.

表 3 三个标记集配对样本 T 检验结果

Sig.(双侧)	标记集 1	标记集 2	标记集 3
标记集 1	1	0.671	0.583
标记集 2	0.671	1	0.711
标记集 3	0.583	0.711	1

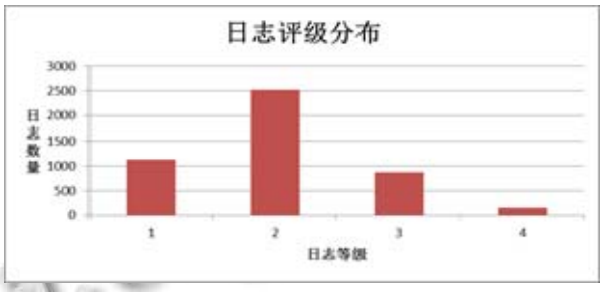

图 6 日志评级分布

4.5 质量特征验证

利用标注的数据对选取的各项质量指标和评分之间关系进行分析, 发现各词性的数量和日志质量之间有较明显的正相关关系, 而中心词词频和日志质量之间有显著的负相关关系. 各相关系数和显著性如表 4 所示. 从表中可以看出, 大部分特征都和日志等级有较高的相关性, 所有特征的相关显著性都小于 0.05, 相关关系明显, 都可作为分类模型的特征.

表 4 日志质量特征和日志质量评分的相关性表

特征	Spearman 相关系数	Spearman 显著性
Word_num	0.847	0.000
WordType_num	0.776	0.000

Verb_num	0.640	0.000
Adj_num	0.282	0.000
String_num	0.595	0.000
Aux_num	0.524	0.000
Con_num	0.415	0.000
Prep_num	0.409	0.000
Verb_user	-0.245	0.000
Verb_all	-0.320	0.000
Noun_user	-0.422	0.000
Noun_all	-0.521	0.000
Similarity	0.153	0.013

4.6 模型训练和效果分析

4.6.1 分类模型选择

使用各种分类模型分别进行训练和建模，综合各项指标，表现最好的是J48决策树算法，决策树模型的结果也比较容易理解和实现. 各种分类模型的效果如表5所示. 可以看出J48和RandomForest的效果最好，通过进一步分析发现RandomForest会将日志质量跨等级误分，而J48只会将相邻两个等级的日志误分，由于跨等级对日志评估影响较大因此选用J48作为日志评估模型.

表5　各分类算法的分类效果

Classifier	accuracy(%)	Kappa statistic	Mean absolute error	Relative absolute error(%)
J48	87.14	0.7899	0.0824	26.79
BayesNET	83.05	0.731	0.0902	29.33
Logistic	86.64	0.782	0.094	30.55
NaiveBayes	77.30	0.647	0.117	38.04
NBTree	85.53	0.769	0.0899	29.24
RandomForest	88.17	0.807	0.0803	26.13

4.6.2 模型效果分析

模型的日志各个质量等级日志的分类准确率如表6所示，第N类就代表第N等级的日志. TP rate最高的是第二类，说明质量等级为二的日志被正确评估出来的比率最高，有90%的都被正确评估. 但同时它也是FP rate最高的，是其他几类日志的数倍，说明被判定为第二等级的日志数量较多，正确评估的覆盖率比较高，同时错误分类到此等级的日志也相对较多. TP rate相对最低的是第四类的0.703，表示有30%的第四等级的日志没有被正确分类，同时它的FP rate最少为0.007，错分为第四类的日志非常少，说明分类器对于第四类日志的预估是比较保守的，较少的日志被分为第四类.

表6　各等级质量日志的分类准确率

Class	TP rate	FP rate	Precision	Recall	F-Measure	ROC Area
1	0.866	0.038	0.88	0.866	0.873	0.96
2	0.902	0.132	0.889	0.902	0.896	0.916
3	0.808	0.04	0.823	0.808	0.815	0.919
4	0.703	0.007	0.783	0.703	0.773	0.965

总体来说，模型的准确率较高，达到87%，且具有较高的可信度，证明本文提出的方法对工作日志报告具有较好的适用性，通过该方法可以合理高效的评估日志的质量，节省了人力物力的同时高效地利用了日志报告提供的信息，为外包单位的监控和决策提供了帮助.

4.6.3 模型展示

本文选择的J48分类算法是决策树算法，模型以树的形式呈现，易于理解和观察，如图7所示. 决策树的每个非叶节点都是一个判断条件，每个叶节点指示一个类别. 从根节点到叶节点是一条判断路径，满足路径上所有条件的样本被分类为叶节点所指类，例如图中标红的一条路径，Word_Num(总词数)少于11且少于4，String_Num(字符串数量)大于零的日志被分类为第二等级的日志. 后面括号内是在测试集中正确分类/错误分类的数量，整个树形结构就是本文的分类评估模型. 分类的结果和日志的等级对应，评估日志质量的规则包含在决策树中. 根据分类模型可以进一步的开发工具，实现对日志质量的自动化解析和评估.

```
Word_Num <= 11
|   Word_Num <= 4
|   |   String_Num <= 0
|   |   |   Word_Num <= 2: 1 (585.0/7.0)
|   |   |   Word_Num > 2
|   |   |   |   Noun_frequency_Discretize <= 5
|   |   |   |   |   Noun_Num <= 1: 1 (55.0/17.0)
|   |   |   |   |   Noun_Num > 1
|   |   |   |   |   |   Noun_frequency_user__Discretize <= 6: 2 (74.0/22.0)
|   |   |   |   |   |   Noun_frequency_user__Discretize > 6: 1 (40.0/18.0)
|   |   |   |   Noun_frequency_Discretize > 5: 1 (388.0/78.0)
|   |   String_Num > 0: 2 (338.0/33.0)
|   Word_Num > 4
|   |   Word_Num <= 6
|   |   |   String_Num <= 0
|   |   |   |   Noun_frequency_Discretize <= 6: 2 (178.0/21.0)
|   |   |   |   Noun_frequency_Discretize > 6
|   |   |   |   |   Similarity <= 0.001745: 2 (156.0/42.0)
|   |   |   |   |   Similarity > 0.001745: 1 (74.0/28.0)
|   |   |   String_Num > 0: 2 (365.0/6.0)
|   |   Word_Num > 6: 2 (1168.0/55.0)
Word_Num > 11
|   Word_Num <= 25
|   |   String_Num <= 0
|   |   |   Word_Num <= 13: 2 (95.0/11.0)
|   |   |   Word_Num > 13: 3 (155.0/40.0)
|   |   String_Num > 0
|   |   |   Noun_Num <= 4
|   |   |   |   Word_Num <= 13
|   |   |   |   |   Noun_frequency_Discretize <= 2: 3 (40.0/18.0)
|   |   |   |   |   Noun_frequency_Discretize > 2: 2 (78.0/24.0)
|   |   |   |   Word_Num > 13: 3 (238.0/44.0)
|   |   |   Noun_Num > 4: 3 (392.0/39.0)
|   Word_Num > 25
|   |   Word_Num <= 32: 3 (135.0/42.0)
|   |   Word_Num > 32: 4 (118.0/9.0)
```

图7　决策树分类模型

5 结语

工作日志是周、月进度报告的补充, 能体现项目更细粒度的进展情况, 对工作日志质量的评估可以为外包项目的监控和项目质量的评估提供有效依据. 为帮助外包项目的监控, 本文提出了一个基于文本分析和机器学习分类算法的人员工作日志质量评估方法, 该方法综合考虑了日志文本的结构、内容、主题等因素, 通过数据分析、特征提取、标注训练集、特征验证、训练模型等步骤建立质量评估模型. 结合一个核高基课题的日志数据进行试验, 建立了一个准确率较高的评估模型, 取得了较好的效果. 外包项目的管理者可以根据本文提出的方法对项目的日志质量进行评估, 为项目的监控提供帮助, 决策提供支持.

本文的数据分析部分中分词使用的词典为通用词典, 对项目中的特殊词汇拆分会有一定的偏差, 下一步计划在分词之前根据日志数据构建分词词典, 增加分词准确性, 进一步提高模型的效果.

参考文献

1 Chrissis MB, Konrad M, Shrum S. CMMI for development: guidelines for process integration and product improvement. Pearson Education, 2011.

2 Simplified Parallel Process and Software Project Management. Applications of the Computer Systems, 2004.

3 胡永祥.软件开发的管理策略.中国包装工业,2002,(5): 132–133.

4 单玥.软件外包项目的运营管理研究[学位论文].北京:北京邮电大学,2013.

5 Dekhtyar A, Hayes JH, Menzies T. Text is software too. MSR 2004: International Workshop on Mining Software Repositories at ICSE'04. Edinburgh, Scotland. 2004. 22.

6 Bettenburg N, Premraj R, Zimmermann T, et al. Extracting structural information from bug reports. Proc. of the 2008 International Working Conference on Mining Software Repositories. ACM. 2008. 27–30.

7 Bacchelli A, Dal Sasso T, D'Ambros M, et al. Content classification of development emails. Proc. of the 2012 International Conference on Software Engineering. IEEE Press. 2012. 375–385.

8 Tomita M. Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems. Springer, 1985.

9 Vaillant P. A chart–parsing algorithm for efficient semantic analysis. Proc. of the 19th International Conference on Computational Linguistics,Volume 1. Association for Computational Linguistics. 2002. 1–7.

10 Che W, Li Z, Liu T. Ltp: A chinese language technology platform. Proc. of the 23rd International Conference on Computational Linguistics: Demonstrations. Association for Computational Linguistics. 2010. 13–16.

11 刘群,张华平,俞鸿魁,等.基于层叠隐马模型的汉语词法分析.计算机研究与发展,2004,41(8):1421–1429.

12 孙昌年,郑诚,夏青松.基于 LDA 的中文文本相似度计算.计算机技术与发展,2013,23(1):217–220.

13 Bouckaert RR, Frank E, Hall M, et al. WEKA Manual for Version 3-7-8. 2013.