软件缺陷数据的分析方法及其实现*)

刘 海 郝克刚
(西北大学信息科学与技术学院 西安 710127)

摘 要 软件缺陷数据的分析对于软件质量保证、项目管理和过程改进具有重要的意义,但目前的缺陷管理工具的数据分析功能普遍比较薄弱。本文首先分析了软件缺陷属性数据的类型,在此基础上阐述了软件缺陷数据分析的基本方法,包括一元数据分析和多元数据分析。最后讨论了实现软件缺陷数据分析方法所涉及到的几个关键技术。

关键词 软件缺陷数据分析,软件缺陷管理,关联规则

Method of Software Defect Data Analysis and its Implementation

LIU Hai HAO Ke-gang
(College of Information Science and Technology, Northwest University, Xi'an 710127, China)

Abstract The analysis of software defect data is important for software quality assurance, project management and processes improvement., but the data analysis functions of current software defect management tools are weak. The data types of software defect attributes are discussed, and then the basic analysis methods of software defect data, including one-variable data analysis and multiple-variable data analysis, are proposed. Moreover, several key technologies for implementing the methods of software defect data analysis are discussed.

Keywords Analysis of software defect data, Software defect management, Association rules

1 引言

在软件项目的开发过程中,会产生大量的缺陷数据,这些数据包含着丰富的有关软件质量、过程、技术和资源等方面的信息,因此对缺陷数据进行分析可以为软件质量保证、项目管理和过程改进决策提供有价值的参考和客观的依据。

目前的软件缺陷管理工具在缺陷数据的分析方面普遍比较薄弱,一般只提供一些缺陷属性的简单统计功能,用户不得不借助一些其它的统计分析软件或自行开发数据分析组件来进行缺陷数据的分析。此外,许多软件组织所采用的缺陷数据分析方法不尽合理,不能充分挖掘缺陷数据所包含的有价值的信息,甚至产生有误导性的分析结果。笔者所参与的"软件缺陷管理和度量系统(BMMS)的开发"项目针对这些问题进行了深入研究,本文将阐述我们在该项目的研发过程中所提出的软件缺陷数据的基本分析方法及其实现技术。

2 缺陷属性值的类型

软件缺陷具有若干属性,属性值的类型可分为文本类型和统计类型。文本类型的属性值是描述性的文本,没有固定的取值范围,例如缺陷的重现步骤和症状描述就属于文本类型的属性;统计类型又分为分类类型(也称标称(nominal)类型)和数值类型,在一般情况下,软件缺陷数据分析的对象是统计类型的属性。

分类类型的属性值取自于一个预先定义好的包含有限个元素的集合,从数据刻度的角度来看,分类类型的属性值属于标称刻度(nominal scale)或序数刻度(ordinal scale)。例如,缺陷的"类别"属性可以取"功能"、"接口"、"算法"等值,"严重程度"属性按照缺陷严重性的高低可以取"关键"、"重要"、"一般"、"细微"等值,它们都是分类类型的属性。此外布尔类型也可以算作一种特殊的分类类型(只取两个值:true或false)。数值类型包括整数、浮点数和日期,属于比率刻度(ratio)或间隔刻度(interval scale),例如修改缺陷所用的时间或成本、发现缺陷的日期等属性都属于数值类型。

在不同的数据刻度上所能执行的运算是不同的,在明确了缺陷属性值的类型和刻度后,就可以针对不同类型的属性采取适当的统计分析方法。

软件缺陷数据通常存放在关系型数据库中,每个缺陷属性对应于表的一个字段。

3 缺陷数据分析的基本方法

软件缺陷数据分析是一个含义广泛的概念,具有数量繁多的分析方法和分析指标,不同的方法和指标面向不同的度量目标,一个缺陷管理工具不可能支持所有的分析方法,但应支持最常用的基本方法。这里所谓的"基本方法"是指仅限于对缺陷本身所包含的属性数据进行分析的方法,不涉及任何其它的度量数据(如软件规模、复杂性、配置管理数据等)。这些方法有的只对单个缺陷属性进行分析,有的对两个以上的缺陷属性进行分析,我们将其分别称之为一元数据分析和多元数据分析。

3.1 一元数据分析

软件组织常常需要了解软件缺陷在某一属性上的分布,例如不同状态、不同严重程度、不同软件模块的缺陷的绝对数量和相对数量,缺陷数量随时间的走势等,这些指标的统计可用一元数据分析来实现。

*)西安市科技计划项目(ZX06028)。**刘 海** 博士研究生,工程师,主要研究方向为软件质量保证和软件测试;**郝克刚** 教授,博士生导师,主要研究方向为软件工程。

对于一个分类类型的属性，一元数据分析是统计各属性值所对应的缺陷个数以及百分比。例如对于缺陷的“严重程度”属性，统计一个缺陷集合中不同严重程度的缺陷的数量和百分比，统计结果可用表格、柱状图或饼状图来表示。

对于数值类型的属性，一元数据分析可计算其均值、总和、最大/最小值、方差、标准差、中位数等，并可绘制出缺陷数量相对于该属性的分布曲线。日期型的属性（如缺陷报告时间和关闭时间）属于间隔刻度，一般只计算时间间隔（如最早的缺陷报告时间与最晚的缺陷报告时间之间的间隔）或绘制缺陷数量随时间的走势图。图 1 为 BMMS 系统中某软件项目下新报告的缺陷数量随日期的走势图。

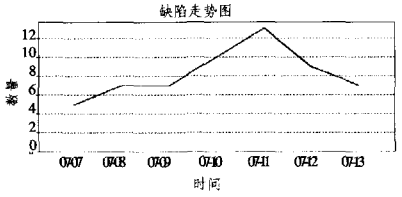

图 1　某项目下新报告缺陷随日期的走势图

一元数据分析得到的信息量是有限的，为了满足项目管理和过程改进的需要，还需使用多元数据分析。

3.2　多元数据分析

多元数据分析用于揭示多个属性的取值分布和属性之间的关联关系。通常多个缺陷属性取值的分布比单个缺陷属性的取值分布更能反映出软件过程和技术中存在的问题，特别是那些频繁出现的属性取值组合对于过程和技术的改进具有很高的参考价值[1,2]。

如果分析对象是分类类型的属性，分析指标包括属性取值组合的数量和百分比以及各属性间的关联规则；如果是多个数值类型的属性，分析指标可以是协方差、散点图和线性相关性；如果是分类类型属性和数值类型属性的混合，一般需要将数值类型属性“离散化”，转换为分类类型后再进行分析，例如修复缺陷所花费的工时（以小时为单位）是一个数值类型的属性，把它按区间划分为 4 种取值，即：少于 1 小时，1 小时到 8 小时，8 小时到 24 小时，24 小时以上，这样就把该属性离散化为一个分类类型的属性。

属性取值组合的数量和百分比可用表格的形式表示，二元取值组合的数量和百分比还可以用图形来表示。图 2 所示为“开发阶段”属性和“缺陷类型”属性的二元分析结果，横轴表示开发阶段，各开发阶段所发现的不同类型缺陷的数量用不同颜色的柱形来表示。

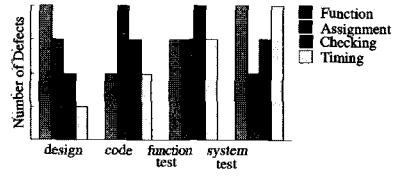

图 2　不同“开发阶段”所发现的各种类型缺陷数量分布图[3]

属性之间的关联性分析属于关联规则挖掘问题。下面先给出几个相关的定义[4]。

定义 1　设 $A=\{A_1,A_2,\cdots,A_m\}$ 是数据集 D 的属性集合，属性 $A_i(1\leqslant i\leqslant m)$ 的属性值集为 R_i，$\langle A_i,u\rangle$ 称为项目（Item），其中 $u\in R_i$；$n(n\geqslant 0)$ 个项目组成的集合 X 称为项目集（Item Set），如果 $|X|=k$，则称为 k-项目集。

定义 2　对于 k-项目集 $X=\{\langle A_1,u_1\rangle,\langle A_2,u_2\rangle,\cdots,\langle A_k,u_k\rangle\}$ 和数据集 D 中的记录 T，T 的 $A_1,A_2,\cdots,A_k$ 属性的属性值分别为 $T_{A1},T_{A2},\cdots,T_{Ak}$，如果对所有的 $i=1,2,\cdots,k$ 都有 $T_{Ai}=u_i$，则称记录 T 包含在 k-项目集 X 中。

定义 3　数据集 D 中包含在项目集 X 的记录数称为项目集 X 的支持数，计为 n，项目集 X 的支持度记作 support (X)，且满足等式：$\text{support}(X)=n/|D|\times 100\%$。

定义 4　若 X,Y 为项目集，且 $X\cap Y=\phi$，蕴涵式 $X\Rightarrow Y$ 称为关联规则，X,Y 分别称为 $X\Rightarrow Y$ 的前提和结论。关联规则 $X\Rightarrow Y$ 的置信度记作 confidence$(X\Rightarrow Y)$，且满足等式：confidence$(X\Rightarrow Y)=\text{support}(X\cup Y)/\text{support}(X)\times 100\%=m/n\times 100\%$，其中 m 为 $X\cup Y$ 的支持数，n 为 X 的支持数。

用简单的语言表达，支持度就是指满足项目集 X 的记录数占总记录数的比例，置信度就是指使关联规则 $X\Rightarrow Y$ 成立的记录数占满足前提 X 的记录数的比例。

关联规则挖掘问题就是在给定的数据集合中产生所有满足指定的最小支持度和最小置信度的关联规则。对于缺陷数据集合来说，使用关联规则挖掘算法可发现缺陷属性之间的关联性[2]。

关联规则挖掘算法有多种，我们采用 Apriori 算法对我们开发的一个软件项目中的缺陷进行了分析，分析的缺陷属性是“活动（Activity）”、“类型（Type）”和“来源（Origin）”，这三个属性的含义和取值如表 1 所示。

表 1　用于关联规则分析的几个缺陷属性

属性名	含义	取值
Activity	缺陷是在执行什么活动时被检测到的。	需求审查（Requirement Inspection）、设计复审（Design Review）、测试（Testing）、用户使用（User Operation）
Type	修复缺陷时对软件所作的修改属于哪一类问题域。	功能（Function）、接口（Interface）、赋值/初始化（Assign/Init）、算法（Arithmetic）、数据结构（Data Structure）、时序（Timing）、用户界面（UI）
Origin	产生此缺陷的原由，最早可追溯到哪个开发活动中的问题。	需求分析（Requirement-Analysis）、设计（Design）、编码（Code）、维护（maintenance）

支持度和置信度分别取 20% 和 90%，分析得到下面 5 个关联规则，每条规则右侧显示出该规则的置信度。

1. Type=Arithmetic ==> Origin=Code conf:(100%)

2. Origin=Requirement-Analysis Activity=Testing ==> Type=Function conf:(91%)

3. Origin=Requirement-Analysis ==> Activity=Testing conf:(90%)

4. Origin=Requirement-Analysis ==> Type=Function conf:(100%)

5. Type=Function ==> Activity=Testing conf:(0.92)

进一步提高支持度，取 30%，置信度仍取 90%，分析得到 2 个关联规则：

1. Type=Arithmetic ==> Origin=Code conf:(100%)

2. Type=Function ==> Activity=Testing conf:(0.92)

从第 1 条规则可以看出，产生了较多的算法缺陷，且都是在编码阶段引入的，由此可以推测：在编码阶段的算法实现技术方面可能存在问题，或者设计阶段的算法设计可能需要进一步改进。第 2 条规则说明绝大多数功能缺陷都是由测试活动发现的，这是一个正常的情况。可见，缺陷数据的关联规则挖掘有可能发现软件过程中潜在的问题，也可以反映开发活

动的效率。

需要说明的是关联关系不等于因果关系，一个关联规则只是说明在前提为真的情况下，结论为真的概率较大。

4 实现技术

在软件缺陷管理工具中，一元和多元数据分析可按以下方式实现：提取出所有统计类型的缺陷属性，将其名称以列表的方式显示在用户界面上，用户可选择一个或多个属性以及对应的分析指标和结果的显示方式，系统根据用户的选择进行相应的分析并显示分析结果。

上述缺陷数据分析方法的实现涉及到以下两个关键技术。

许多分析结果需要以柱状图、饼图和线图等可视化的形式呈现给用户，虽然可使用开发平台提供的底层图形 API(如 Java 的 2D 和 3D 图形 API)来实现分析结果的可视化，但实现方法较为复杂，且需要开发者熟练掌握图形编程技术。更为常用的方法是使用现有的图形软件包如 JFreeChart，WebChart3D 等来实现图形输出，这些软件包提供了丰富的统计图形绘制功能，可根据分析结果在普通用户界面或 Web 页面上生成各种图形。

缺陷属性关联规则的分析需要调用关联规则挖掘算法，实现这种算法要求具有一定的统计分析和数据挖掘知识，但开发者也可以借助一些已有的工具来实现该算法。例如，新西兰怀卡托大学开发的“Weka”就是一个功能全面的数据挖掘工具，它用 Java 语言实现了几乎所有常用的数据挖掘算法，而且提供了调用各种算法的 API，使开发者可以将该工具中的算法嵌入到自己的应用程序中[5]。

结束语 除本文所介绍的方法以外，软件缺陷数据分析还有许多分析指标和分析方法，如软件可靠性模型的建立，使用线性回归和贝叶斯网等技术预测缺陷数量等。实际上，有关软件开发过程、技术和资源的许多问题都与软件缺陷密不可分，软件组织在项目进展过程中可能会提出各种各样的缺陷度量要求，缺陷数据的分析也常常与缺陷之外的度量数据相联系，因此缺陷分析工具应具有很好的可扩展性标准的数据接口，以便于根据需要增加新的数据分析功能。在今后的工作中，我们将更为全面地研究缺陷分析方法，并开发功能更强、更为简便实用的缺陷分析工具。

参 考 文 献

[1] Dalal S, Hamada M, Matthews P, et al. Using Defect Patterns to Uncover Opportunities for Improvement // Proceedings of International Conference on Applications of Software Measurement (ASM). San Jose, CA, , February 1999: 15-19

[2] Menzies T, Lutz R, Mikulski C. Better Analysis of Defect Data at NASA // Proceedings of SEKE2003, the Fifteenth International Conference on Software Engineering and Knowledge Engineering. San Francisco, July 2003

[3] Bridge N, Miller C. Orthogonal Defect Classification-Using Defect Data to Improve Software Development. Software Quality, 1998, 3(1): 1-8

[4] 毛国君，段丽娟，王实，等. 数据挖掘原理与算法. 北京：清华大学出版社，2005

[5] Witten I H, Frank E. 数据挖掘-实用机器学习技术(英文版). 第2版. 北京：机械工业出版社，2005

(上接第 228 页)

确率。只要说明不同的语义图像簇其聚类的底层特征空间是不同的，即不同图像簇中各个特征的影响因子是不同的，前人通过相关反馈来修改特征权重以及我们此处的实验都证明了这一点。此处改进的聚类检索使得类内相似度高，类间相异度高，达到良好的聚类效果，从而 IC-CBIR 就能够取得较好的检索准确率。

表 1 特征相似度平均值

类别 \ 相似度及权重	颜色 C	文理 T	形壮 S
类 1 鸟(100)	0.387 (25.8%)	0.701 (46.7%)	0.412 (27.5%)
类 2 蓝天海洋(100)	0.754 (46.6%)	0.548 (33.9%)	0.316 (19.5%)
类 3 建筑(100)	0.522 (31.7%)	0.649 (39.4%)	0.475 (28.9%)

表 2 聚类结果分析

类别 \ 相似度及权重	图像数目	普通聚类匹配数目	改进聚类匹配数目
簇 1	100	67	83
簇 2	100	74	69
簇 3	100	59	72

本文以从 http://www.wang.st.psu.edu/IMAGE 下载的包含 10,000 幅测试图像的图像库为实验素材。其中图片规格为 128 * 85 或 85 * 128。首先从图像库中人为地选出 3 类图像，每一类 100 张。为了减弱个人的主观因素，采用多人选取取交集的办法。然后计算类内图像在颜色、纹理、形状三个特征上的相似度。这里类的特征相似度是通过类内所有图像特征相似度的平均值，其结果见表 1，聚类的结果分析见表 2。

由表 1 可以看到，在不同的语义簇内，各个特征所占的比重不是同样大的，这就说明了提取关键特征空间、计算簇特征影响因子的必要性。虽然初次聚类形成簇的特征影响因子与语义上完全正确的簇的簇内特征影响因子并不完全相同，但它们大致接近。由表 2 的聚类准确率可以看到，二次聚类的准确率为 74.7%，而按照统一特征空间的普通聚类的准确率只有 66.7%。这样，新簇的簇内图像语义上更加接近，图像的检索效果也就更好。而且改进的聚类使得一幅图像可以同时属于两个或是多个簇，更加符合语义的模糊性。

结束语 随着多媒体技术、网络技术的发展和需求的带动下，对于它的研究方兴未艾。本文旨在缩小或解决语义鸿沟，提出了一种新颖的 CBIR 检索模型，将语义关键字与底层视觉特征相结合来进行检索。而语义关键字的产生是由用户的相关反馈得来的，具有较高的准确性。这样，用户对图像库检索的次数越多，检索准确率和检索速度也会越好。

参 考 文 献

[1] 温超，耿国华. 基于内容图像检索中的“语义鸿沟”问题. 西北大学学报(自然科学版)，2005

[2] 夏定元. 基于内容的图像检索通用技术研究及应用. 博士学位论文. 2004

[3] 张莉，周伟达，焦李成. 核聚类算法. 计算机学报，2002

[4] 黄元元. 基于视觉特征的图像检索技术研究. 博士学位论文. 2003