信用构件的刻面分类及检索方法研究①

Research on Facet Classification and Retrieval Method of Credit Component

范　菁　刘　韬　熊丽荣　（浙江工业大学软件学院 浙江杭州 310023）

摘　要: 本文引入软件复用来实现信用评估系统的构件化开发,通过对信用评估系统及评估建模方法的分析,提出信用评估构件库的刻面分类,结合具有良好扩展性的术语空间层次编码,对构件检索技术进行了讨论,将查询条件和构件描述的匹配转化为字符串集之间的匹配,采用向量空间模型的相似度量来提高构件的查全率。

关键词: 构件库 刻面分类 检索 软件复用 信用评估

1　引言

引入软件复用思想来实现信用评估系统的构件化开发与组装,可以充分利用已有的高质量软件资产,不仅能提高开发效率,而且在系统的稳定性、扩展性、可重构性上也得到大幅度提升。

存在大量可复用信用构件是复用的前提,构件库是支持信用评估系统实现构件化开发的一个重要基础设施,其中构件的表示和检索技术是可复用软件构件库两个主要核心技术[1]。目前,刻面的表示方法以其较强的表示能力和灵活的扩展机制已经成为构件库系统主要采用的构件表示方法,在此技术上的构件检索技术已得到软件复用界的重视和研究。

刻面描述的构件检索主要以传统的数据库检索技术为主,并结合利用同义词词典和刻面术语空间的层次结构来实现构件的检索[2]。由于数据库表的多次关联会使检索效率较低,构件检索也要求对查询条件有一定的模糊匹配能力,保证一定查准率的情况下提高查全率,并返回相应的匹配程度作为参考。本文提出了信用领域的构件刻面分类模式,运用层次编码技术,将检索转化为字符串集之间的匹配,并通过进行了向量空间模型计算相似度量。

2　刻面分类

刻面[3]术语最早使用在50年代,分别出现在Ranganathan的图书馆分类系统中和Guttman的一项社会调查中。刻面分类方法具有较高的扩展性和灵活性,它通过从刻面表中选择定义好的词汇,然后合成这些词汇形成合成的分类。1987年,Prieto – Diaz和Freeman系统地提出了用刻面分类方法来对可复用软件构件进行分类与组织的思想[4],这种方式通过反映构件本质特性的视角(刻面)对构件进行精确的分类。一个刻面分类模式(Faceted Scheme)由一组描述构件本质特征的刻面组成,每个刻面从不同的侧面对构件库中的构件进行分类。每个刻面有一组术语(Term),术语之间具有一般/特殊关系而形成结构化的术语空间(Term Space),允许术语之间有同义词关系,术语空间可以演变。

构件的刻面表示方法与其它构件表示方法相比,具有以下两个优点[5]:(1)对构件进行了多视角下的分类描述;(2)构件分类描述的术语空间易于修改与维护。由于其良好的扩展性、灵活性以及易理解性,在构件的表示上立刻得到了普遍的研究与应用,例如:REBOOT构件库中定义了4个刻面:抽象、操作、操作对象和依赖;我国北京大学的青鸟构件库定义的5个刻面为:使用环境、应用领域、功能、抽象层次和表示方法。

3　信用构件库的刻面分类

3.1　信用评估系统

信用评估是对客户偿还债务能力和意愿的评估,

① 基金项目:浙江省科技计划面上项目《基于 Web Service 封装的信用构件库》(2007C21011)

是对信用风险的综合评价,以评价结果来减小违约风险,实现信用制度的规范化管理。

信用评估系统首先获取被评估对象的特征信息,如企业基本信息、财务报表、行业特征等反映信用状态的数据,通过对这些特征信息的提取和计算,生成一些信用评估模型的指标变量,最后代入到信用评估模型计算信用值,以一定的形式反馈被评估对象的信用情况。信用评估体系结构如图 1。

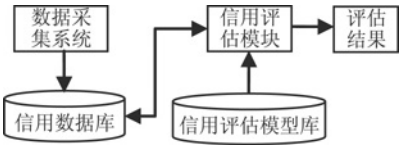

图 1 信用评估系统结构

3.2 评估模型的建立

随着市场经济的深入发展,企业对信用评估有迫切的需求。当信用模型库没有合适的信用评估模型时,可利用模型构造工具来重新建立新的评估模型,以最合适的信用评估模型来对特定的客户进行评估。

建立企业信用评估模型的关键有两点:一是建立一个全面的、准确的、客观的企业信用评估指标体系;二是选用适当的方法建立信用评估模型。指标体系中各指标有不同的量纲,给评估带来许多困难,需要将不同量纲的评价指标,通过适当的变换化为无量纲的标准化指标,使评价指标标准化。定量指标一般有线型比例变换、极差变换、向量归一化、功效系数法,定性指标有三角模糊数、专家打分法等。

确定了评估指标之后,选择适当的模型方法进行评估模型的建立,目前穆迪等评估机构根据建立评估模型的方法不同,分为五种类型:传统专家方法、财务比率综合分析法、基于统计的方法、基于金融理论的方法和基于现代信息技术的方法。

完成模型的制定以后,为了衡量模型预测能力的强弱,必须进行模型检验[6]。主要用到的模型检验报告有:交换曲线、K – S 指标、区分度、拟合度曲线等。

3.3 信用构件库的刻面分类

基于构件的软件开发中,有两个视角通常被区分开,一个是构件的开发者的视角,一个是构件的使用者视角[7]。开发者使用传统的软件开发技术开发构件实体,使用者使用的是基于构件的软件开发。

构件的描述结构是构件存储和检索的基础,可以用诸多的信息来表示,有构件名称,接口,开发环境,应用环境,制作者,构件评价和可靠性等静态和动态两个方面。

刻面的表示方法主要是从使用者的视角来描述,并且往往从构件的功能和应用环境等静态角度描述构件库。因此本文选择最能反映和区分构件本质的属性作为刻面,并建立使用者在查询信用构件时最为关注的属性的刻面术语空间。

通过上述对信用评估领域的分析,本文选定以下四个构件属性作为信用评估构件库的刻面:

(1) 构件类型:构件类型为构件在其制作过程中所遵循的工业标准。不同的复用环境选择不同的构件类型。

(2) 功能:该构件在原有或可能的软件系统中所提供的软件功能集,任何构件都必须提供至少一种功能。

(3) 应用领域:构件被应用于信用评估系统的具体子模块。

(4) 使用环境:使用该构件时必须提供的硬件和软件平台,如特定的操作系统,数据库平台,容器等。

以上四个属性彼此之间相互正交,充分体现构件对于用户最相关的特征表示,能较好适应信用评估构件库的发展和刻面的兼容扩充。该刻面分类建立的术语空间如下:

(1) 构件类型术语空间

主要有 JavaBeans、EJB、Java Applet、Java Servlet、AWT、Web Service、COM/DCOM、CORBA 等。

(2) 功能术语空间

可分为数据管理、统计描述、相关性分析、建模、检验、图表显示、流程相关等。数据管理又分用户管理、元数据管理、数据采集、数据抽取、转换、装载、数据存储、发布;统计描述又分均值、中位数、众数、全距、方差、标准差、四分为数等;相关性分析有主成份分析、因子分析、变量类聚分析;建模有统计学方法,包括判别分析、Logit 分析、Probit 分析、递归分类树、最邻近方法,非统计学方法有神经网络模型、模糊神经网络方法,现代风险度量模型有 CreditMetrics、KMV、CreditRisk + ;检验有交换曲线、区分度曲线、ROC 曲线 Gini 系数、

拟合度曲线、区分统计量 ;图表显示包括表格、柱状图、折线图、饼图、面积图。

(3) 应用领域术语空间

应用领域术语空间有信用评估模型,包含了定量、定性、综合、古典评估模型、自定义评估模型 ;信用管理流程包括业务管理、资金管理 ;信用数据管理 ;建模工具等。

(4) 使用环境术语空间

使用环境术语空间包括了容器、操作系统、数据库。运行容器有 Tomcat、Jboss、WebLogic 等。操作系统分为 Windows、UNIX 等。数据库有 SQL Server、MYSQL 等。

随着评估领域的细化和深入,在对构件的描述过程中,根据需要逐步的修正。添加新术语或将术语的同义词加入同义词库中,对一些不能准确表达构件属性的术语进行剪除或修改,力求构件的制作者与使用者在构件描述文档的理解上的一致。

4 信用构件的刻面检索

领域内积累大量的信用领域基础和专用构件后,如何对大量的构件进行有效的检索是构件能否成功复用的关键。对于信用构件的刻面分类方案,将其中的刻面、子刻面分别映射为树中对应的父节点、子节点,用一个虚拟根节点组合为刻面术语空间树。信用构件的刻面表示将从术语空间树选择一定的术语进行描述,形成描述树。

刻面术语空间是一个可拓广、拓深的树形层次结构,对其数据进行检索时,需要多次遍历才能判断两个术语是否存在父子关系或祖先关系,时间复杂度极高。采用具有良好扩展性和蕴含路径信息的层次编码[8],可避免数据库表的多次关联,并支持与查询条件的模糊匹配。

4.1 术语编码

针对树形结构的术语空间,可以对术语进行层次编码。具体编码规则为 :每个术语节点对应一个唯一编码 ;编码形式为 $N_1/N_2/N_3/\cdots/N_L$,由 L 个域组成,L 表示该节点所在的层次号,每个域的取值都是整数,表示在当前父节点下的相对位置,取值范围为 0,1,2,……。每层之间用分隔符‘ / ’连接。编码示意如图 2,虚拟根节点术语编码为 ROOT。

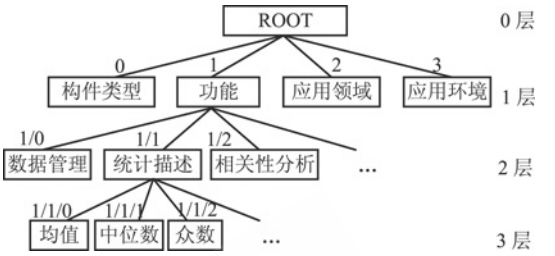

图 2　信用评估构件刻面术语空间编码示例

采用由父节点编码连接当前兄弟节点的排序作为当前节点的编码,术语空间可以无限的扩展,表示足够的信息,有利于叶子节点很多的构件刻面分类的编码。根据上述规则,可得出术语节点间有下列上下位关系:

以 C(Vi) 表示 Vi 节点的术语编码,tc 表示 C(Vi) 的最后域的值,令符号∅表示 tc 等于 C(Vj) 对应域

(1)

若 C(Vi) 是 C(Vj) 前缀,Vj 的域数比 Vi 大 1,表示节点 Vi 是 Vj 的父术语。

(2)

若 C(Vi) 是 C(Vj) 前缀,Vj 的域数比 Vi 大于 1,表示节点 Vi 是 Vj 的祖先术语。

对于图 2,C(功能) =1,C(均值) =1/1/0, C(功能)为 C(均值) 的前缀,L(均值) - L(功能) =2,所以功能为均值的祖先节点。同理,统计描述为中位数的父节点。

4.2 刻面检索

首先通过刻面的树形结构选择最能反映构件特征的叶子节点对信用构件进行描述,叶子编码隐含了其父术语和祖先术语的信息。根据描述术语生成用于构件库互操作的构件描述 XML 文档,将各叶子编码排序后用‘ # ’连接,字符串信息存入数据库构件码表。

系统通过查询接口,将刻面的检索条件转化成一个基于术语空间编码的查询树,与构件描述树实现树与树的匹配,术语之间关系转化为字符串之间的关系,刻面检索流程如图 3。用户通过刻面的树形结构表示选择查询术语集,每个基准术语的有其同义词显示,将选择术语构造为查询树,设置各个术语查询权重。将表示查询树的字符串与构件树进行匹配,返回匹配结果集,根据需要选择是否需要二次检索。

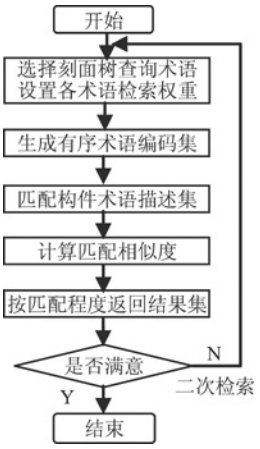

图 3 刻面检索流程

刻面检索转化为输入的查询树 Q 的各节点编码与构件描述术语编码的直接匹配,避免了对数据库的递归查询,具体的匹配算法如下:

输入:查询术语集 Q;构件描述术语集 D。

输出:按匹配度排序的构件集合 R。

1)选择刻面树术语,生成按字母升序排列的字符串编码集;

2)R 为空集;

3)for D 中的每构件术语集 d

4)for(i=0,j=0;小于查询术语数并且 j 小于构件描述术语数)

5)if(查询术语编码 Cq 与描述术语编码 Cd 相等){

6)i++;j++;}

7)else if(查询术语为描述术语祖先节点){

8)i++;}

9)else if(对应域 Cq 小于 Cd){

10)i++;}

11)else if(对应域 Cq 大于 Cd){

12)j++;}

13)end for;

14)根据检索术语权重和向量空间模型计算相似度;

15)end for;

16)将相似度大于阈值的结果返回;

在匹配过程中,以查询术语编码至少为描述术语编码前缀即为匹配。按向量空间模型计算相似度,按匹配度返回结果,保证了松弛匹配,又提高了查全率。

4.3 检索相似度

构件检索时的相似度反映的是用户提交的检索条件与构件描述之间的相似程度,构件的描述采用的是表现构件特征的关键词通过正交原则组织成构件的刻面信息,系统通过刻面的树形结构表示选择构件的描述和查询术语集,因此向量空间模型[9]的内积计算和余弦计算可有效反映匹配相似度,本文采用内积计算法。构件描述定义如下:

定义 1:

* T_i 为构件在第 i 刻面下的术语集;
* ω_i 为第 i 刻面下的术语权重,同一刻面下的术语具有相同权重;
* F 为刻面总数。

刻面检索时用户提交刻面术语空间下的术语集,可用 Q 表示为:

定义 2:

* T_i' 为提交的在第 i 刻面的检索术语集;
* F 为刻面总数。

定义 3:

* m 表示检索是在第 i 刻面下提交了 m 个术语;
* 术语对应权重为 $\omega_{i,M}'$,由用户检索时设置。

构件 D_n 与检索条件 Q 相似性计算为:

定义 4:

如果 T_i 和 T_i' 正交,则值为 0,即提交的检索术语集 Q 在构件描述术语集 D_n 的 T_i 下,没有任何术语被匹配时,相似值为 0;

如果 T_i 和 T_i' 不正交,即提交的检索术语集 Q 在构件描述术语集 D_n 的 T_i 下,至少有一个术语被匹配,相似值为被匹配术语的检索权重值与刻面术语权重值乘积的总和。

定义 5:

MAX(SIM(D_n,Q))为在所有查询术语都能被匹配情况下计算出来的相似值。

计算完成用户查询式和构件描述的相似度后,还需对其结果进行过滤,过滤的效率应以查准率和查全

率来衡量。设置过滤查询结果的阈值是查准率和查全率效果综合相对较好的值,可根据特定实验数据经验获取,也可采用相对更为精确的基于 Boosting 机制计算的相似度阈值。

检索过程以 Logit 模型构件为例,描述的基准术语有 EJB、Logit 分析、Logit 模型、Jboss4. 2. 1GA、windowXP、SqlServer2000,默认权重为 1,按层次编码规则,其描述编码为:

0/1、1/3/0/1、2/1/3/0/0、3/0/1、3/1/0、3/2/0

用户通过刻面术语显示树选择查询术语并设置查询权重:EJB(2)、Logit 模型(3)、MySQL(1),查询编码为:

0/1、2/1/3/0/0、3/2/1

计算 SIM(D, Q) =1 * 2 +1 * 3 +1 * 0 =5;

MAX(SIM(D,Q)) =1 * 2 +1 * 3 +1 * 1 =6;

所以相似度 =5/6 =0. 833。

5 总结

刻面检索是检索代价、复杂性和检索质量三者最为均衡的方法,适合大规模的构件管理。结合术语空间的编码和相似度计算提高查全率,并以 Lucene 全文检索、属性/值和构件间的关系导航等多种方式辅助查询,成为目前构件复用中检索技术中的一个良好的解决方案。随着构件库检索技术的发展,本文的下一步工作是建立信用领域本体使检索更具联想性和扩展性,并提供统一接口实现构件库的跨库查询。

参考文献

1 王渊峰, 薛云皎, 张涌, 朱三元, 钱乐秋. 刻面分类构件的匹配模型. 软件学报, 2003, 14(3): 403 -405.

2 Ruben Prieto - Diaz. Implementing faceted classification for software reuse. Communications of the ACM, 1991, 34(5): 88 -97.

3 Uta Priss. Faceted Information Representation. 8th International Conference on Conceptual Structures Logical, Linguistic, and Computational Issues, August 2000.

4 Ruben Prieto - Diaz, Peter Freeman. Classifying software for reusability. IEEE Software, 1987, 4(1): 6 -16.

5 Sorumgard, L. S, Sindre, G, Stokke, F. Experiences from Application of a Faceted Classification Scheme. In: Proc. Reuse 93, Lucca, Italia: IEEE CS Press. March, 1993.

6 陈建. 信用评分模型技术与应用. 北京: 中国财政经济出版社, 2005.

7 Casanova M., Van Der Straeten, R., Jonckers V. Supporting Evolution in Component - Based Development Using Component Libraries. In: Proceedings of the Seventh European Conference on Software Maintenance and Reengineering. Washington, DC, USA: IEEE Computer Society, 2003. 123 -132.

8 袁军鹏, 陈铿, 黄进, 李连宏. 一种新的通用概念层次编码方法. 计算机工程, 2004, 32(12).

9 Salton G, Wong A, Yang C S. A vector space model for automatic indexing. Communications of ACM, 1975, 18(11): 613 -620.