基于分级策略和聚类索引树的构件检索方法

王文霞
(运城学院 计算机科学与技术系,山西 运城 044000)

摘 要:基于刻面的构件表示法,其术语空间需要人工建立和维护,具有较强的人为主观性。针对此问题,文中采用刻面分类与全文检索相结合的构件表示方法,提出了一种基于分级策略和聚类索引树的构件检索方法。该方法采用基于语义相似度与优化的构件聚类算法构建构件聚类索引树,并为每个刻面引入合理的权重因子。在真实构件库上的实验结果表明:基于分级策略和聚类索引树的构件检索方法是有效的,相比没有引入分级策略的构件检索方法具有较高的构件查全率和查准率。

关键词:刻面分类;聚类分析;语义分析;索引树;分级策略
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2016)04-0110-04
doi:10.3969/j.issn.1673-629X.2016.04.024

A Component Retrieval Method Based on Classified Policy and Cluster Index Tree

WANG Wen-xia
(Department of Computer Science and Technology, Yuncheng University, Yuncheng 044000, China)

Abstract: In the component representation method based on faceted classification, the term-space needs human to build and maintain so as to have strong subjectivity. Therefore, a component representation method combined faceted classification with full-text retrieval has been adopted in this paper. Meanwhile, a component retrieval method based on classified policy and cluster index tree has been proposed. In this method, the component cluster index tree is built by use of a component clustering algorithm based on semantic similarity and optimization, and the reasonable weight factors are introduced for each facet. On the foundation of the real component library, the experiment shows that by the comparison of the component retrieval method without weigh factors, the component retrieval method proposed has higher precision ratio and recall ratio, and to some extent, achieves component semantic retrieval.

Key words: faceted classification; cluster analysis; semantic analysis; index tree; classified policy

0 引 言

软件复用是提高软件生产率和质量的有效途径,其核心是软件构件技术。而软构件技术领域中,构件的分类与构件的检索是亟待解决的两大主要问题[1-4]。构件分类的合理性是实现构件高效检索的有效途径和关键因素,高效的构件检索可以降低构件理解和查询的成本[5-9]。因此,合理有效的构件分类和准确高效的构件检索,将会大大促进软件的复用,进而促进软件产业的快速发展。

现有的构件分类有多种。W. Frakes 将基于构件的表示划分为信息科学方法、超文本方法和人工智能方法。其中,信息科学方法具有对构件多视角分类描述的特点,在实际中得到广泛应用。刻面分类法属于信息科学方法中的一种,其基本思想是将反映构件本质特性的各个刻面及相关术语置于一定上下文中,实现对构件的精确分类。该表示法可表达丰富的构件信息,是检索代价、检索质量和复杂性三者较均衡的方法,适合于大规模构件库的管理[10];但是,刻面分类方法中构件表示所依赖的术语空间需要人工建立和维护,带有较强的人为主观因素,其结果可能导致用户无法检索到真正所需的构件[11]。

针对此问题,文中基于刻面分类的构件表示法和全文检索方法结合的方法描述构件,以达到降低刻面

收稿日期:2015-09-05 **修回日期**:2015-12-10 **网络出版时间**:2016-03-22
基金项目:国家自然科学基金资助项目(11241005);山西省高等学校教学改革研究项目(J2012098);运城学院教学改革研究项目(JG201418)
作者简介:王文霞(1979-),女,讲师,硕士,研究方向为信息检索、数据挖掘、算法分析与研究。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160322.1522.102.html

表示的主观性问题;同时采用基于语义相似度与优化的构件聚类算法构造构件聚类索引树,并引入分级策略,提出一种基于分级策略和构件聚类索引树的构件检索方法,实现对构件更准确、更高效的检索。

1　相关概念

1.1　构件的分类表示

文中对构件的描述采用刻面分类表示法与全文检索相结合的方法。该方法首先依据某种刻面分类方案(如以刻面的完整性和独立性定义的刻面分类方案——功能、操作对象、使用环境、构件形态和表示及性能共 5 个刻面[12])获取构件的描述文本相对应的每个刻面值;然后采用全文检索的方法对每个刻面下的构件进行聚类分析获取构件的分类描述。这种构件表示法不仅实现了刻面值由受控词到文本的转变,减少了术语空间需人工建立和维护而存在的主观因素,而且构件文本经过刻面分类后内容更为集中,更有利于构件相似度的计算,从而提高基于全文检索的构件聚类精度,实现对构件更合理的分类描述。该构件分类表示法如图 1 所示。

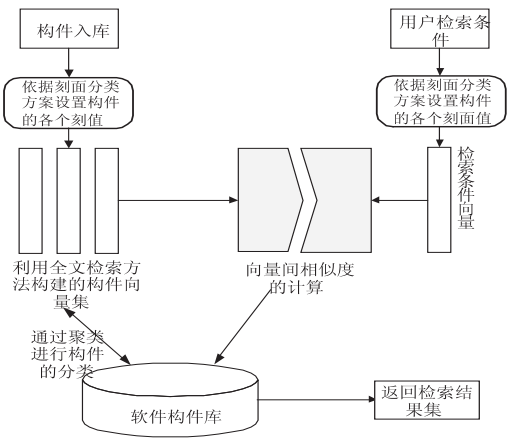

图 1　构件分类表示

1.2　构件聚类索引树

构件聚类索引树(Component Cluster Index Tree, CCIT)是一棵非空的 4 层结构的聚类索引树[13],它满足:

(1)有且仅有一个根节点(root),代表构件库中的所有构件。

(2)父节点中包含着指向子节点的指针和子节点的信息。

(3)叶节点中包含着指向某个具体构件的指针。

(4)第一层为根节点,第二层为刻面层,第三层为类层,第四层为构件层。类层包含着指向该类中所有构件(叶节点)的指针,同时包含着类特征词信息;构件层即叶节点层除了包含条件(3)中的内容,还带有构件特征词信息。其结构如图 2 所示。

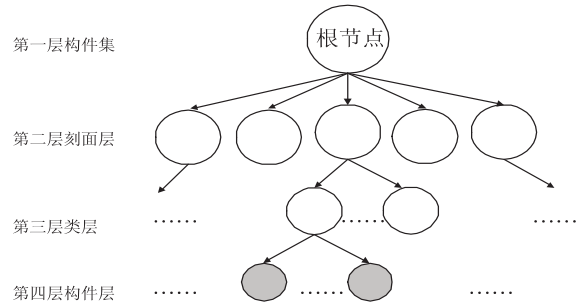

图 2　构件聚类索引树

该聚类索引树构建的基本思想是首先基于刻面分类方案对所有构件进行初次分类,形成索引树的第二层;然后针对每个刻面下的初次分类结果,采用相应的聚类算法进行聚类分析,形成索引树的第三层(类层)和第四层(构件层)。文中第二层采用了基于上述刻面分类与全文检索相结合的构件分类表示方法对构件进行初次分类;索引树的第三、四层采用了基于语义相似度与优化的构件聚类算法实现对构件的分类。其基本思想是:首先基于知网的语义相似度计算方法从语义角度获取构件文本间的相似度;再采用最近邻聚类和遗传算法相结合的方法实现对构件的优化聚类分析[14]。

2　构件检索方法

基于分级策略和聚类索引树的检索方法的基本思想是:以基于语义的构件聚类索引树为基础,依次计算出不同刻面下检索条件与类层中各节点的相似度;然后计算检索条件与不同刻面下相似度最高的类中的各个构件的相似度;接着引入分级策略,为刻面层的各个节点(即不同刻面)设置不同的等级权重值,进而计算其不同刻面与检索条件的最终相似度值,并获得相应刻面下相似度较高的构件集合;再次,对各个刻面下所求的构件求交集,并对不同的刻面下同一构件的相似度求和,求其检索条件与某个构件的总相似度值;最后,依据总相似度值,对获取的构件进行排序,进而便于用户获取所需构件。

其中,(1)检索条件与类层次节点间的相似度采用如下文本相似度计算公式:

$$\mathrm{Sim}(D,D_i)=\cos\frac{\sum_{k=1}^{n}W_{Dk}*W_{D_ik}}{\sqrt{\sum_{k=1}^{n}W_{Dk}^2\sum_{k=1}^{n}W_{D_ik}^2}} \tag{1}$$

式中:W_{Dk} 表示特征词 k 在检索文本中的权重值;W_{D_ik} 表示特征词 k 在第 i 个类文本中的权重值。

(2)检索条件与构件库中第 i 个刻面下第 j 个构件的相似度计算公式为:

$$S_{ij}=\alpha_i(S_{im}+S_{mj}) \tag{2}$$

式中：α_i 表示刻面层第 i 个节点的权重值，且 $\sum_i \alpha_i = 1$；S_{im} 表示在第 i 个刻面下与检索条件相似度最高的 m 类之间的相似度值；S_{mj} 表示与相似度最高的 m 类中第 j 构件的相似度值。

(3)检索条件与构件的总相似度计算公式为：

$$S_j = \sum_{i=1}^{\text{刻面数}} S_{ij} \tag{3}$$

基于分级策略的构件检索算法如下：

输入：所要检索的构件文本；

输出：相似度从大到小的 N 个构件。

Step1：提取检索构件文本的特征词，形成检索特征词向量；

Step2：$i = 1$；

Step3：采用式(1)，计算刻面 i 下检索特征词向量与每个类特征向量的相似度；

Step4：对 Step3 所得到的结果进行排序，获得刻面 i 下相似度较高的前 m 个类；

Step5：采用式(2)，计算检索特征词向量与刻面 i 下前 m 个类中每个构件的相似度值 S_{ij}；

Step6：对 S_{ij} 进行排序，获取刻面 i 下相似度较高的前 P 个构件；

Step7：i ++；转向 Step3，直至 i >刻面数；

Step8：对每个刻面下所取得的前 P 个构件求交集，获取与检索向量较高的构件集；

Step9：采用式(3)，求得检索向量与所得的构件集中的最终相似度值；

Step10：排序，获取最终相似度较高的前 N 个构件，返回。

3 实验结果及分析

文中基于 Matlab7 仿真平台和 Eclipse 开发环境对算法进行了实现；同时与没有引入分级策略的构件检索方法进行比较，来验证基于分级策略和构件聚类索引树的检索方法的有效性。

实验数据来自于上海构件库的构件数据，它包含了六个主题：加密解密(142 个)、文本处理(213 个)、编译器(279 个)、图像处理(362 个)以及数据转换(42 个)和防火墙(102 个)；并采用刻面分类与全文检索相结合的方法描述构件。其中，分级策略中为刻面层上各个节点权重值的设定是通过多次实验并结合用户的关注点(通常，用户比较关心构件的功能刻面)来获取的，这 5 个刻面的分级权重值分别为：$\alpha_1 = 0.423$，$\alpha_2 = 0.218$，$\alpha_3 = 0.097$，$\alpha_4 = 0.125$，$\alpha_5 = 0.137$。

对于实验结果的评价采用查准率和查全率两个指标，其定义如下：

$$\text{查准率} = (N_s + N_a)/M \tag{4}$$

$$\text{查全率} = (N_s + N_a)/N \tag{5}$$

式中：N_s 表示构件检索结果中相似的构件数量；N_a 表示构件检索结果中正确的构件数量；M 表示构件检索结果的构件总数量；N 表示构件库中所有相似构件的总数量。

表 1 给出了三组数据情况下两种算法的查全率和查准率。图 3 给出了两种方法下查准率和查全率的折线对比图。

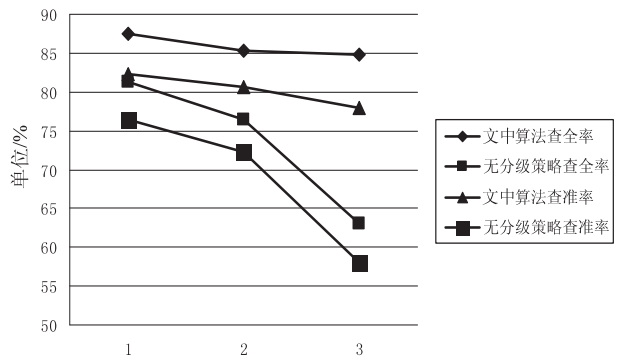

图 3 实验结果对比

表 1 实验结果

加密解密	文本处理	编译器	图像处理	数据转换	防火墙	N_s	N_a	M	N	查全率/%	查准率/%	说明
47	71	93	120	42	34	11	3	17	16	87.5	82.30	文中算法
						11	2	17	16	81.25	76.40	无分级策略
94	142	186	240	42	68	25	4	36	34	85.29	80.56	文中算法
						24	2	36	34	76.47	72.22	无分级策略
142	213	279	362	42	102	33	6	50	46	84.79	78.00	文中算法
						26	3	50	46	63.04	58.00	无分级策略

从表 1 的三组实验结果可以看出，基于分级策略和聚类索引树的构件检索方法其查全率均达到了 80% 以上，查准率也基本在 80% 左右，与无分级策略的构件检索方法相比，文中方法提高了构件检索的查全率和查准率，即提高了检索质量，从而验证了该算法的有效性。同时，通过图 3 中两种方法查全率和查准

率的折线对比,可以看出随着数据量的不断增加,文中算法基本处于平稳的变化中,幅度在 5% 以内;而无分级策略的构件检索方法在第三组数据的测试中,查全率和查准率均出现了急剧的降低趋势,降低幅度达到了 10%,从而表明了文中算法是较为稳定的。

4　结束语

文中为克服刻面分类所存在的主观因素,采用了基于刻面分类与全文检索相结合的构件表示方法以及分级策略,提出了一种基于分级策略和聚类索引树的构件检索方法。该方法具有较高的构件查全率和查准率,而且具有稳定性,可以避免刻面分类的主观性,便于普通用户的查询。但是,该构件检索方法依赖于语义分析技术的发展,因此,将会在不断发展的语义分析技术的基础上,对基于语义的构件检索进一步进行改进和完善,进而更好地满足用户的检索需求。

参考文献:

[1] 王渊峰,张　涌,任洪敏,等. 基于刻面描述的构件检索[J]. 软件学报,2002,13(8):1546-1551.

[2] Mili H, Mili F, Mili A. Reusing software: issues and research directions[J]. IEEE Transactions on Software Engineering, 1995,21(6):528-562.

[3] Rine D C, Sonnemann R M. Investments in reusable software: a study of software reuse investment success factors[J]. Journal of System and Software,1998,41(1):17-32.

[4] 杨芙清. 软件复用及相关技术[J]. 计算机科学,1999,26(5):1-4.

[5] 常继传,郭立峰,马　黎. 可复用软件构件的表示和检索[J]. 计算机科学,1999,26(5):45-49.

[6] 姚全珠,丁新村,冉占军. 基于 XML 的树匹配构件检索算法的研究与实现[J]. 计算机应用研究,2008,25(4):1013-1015.

[7] Emmerich W, Kaveh N. Component technologies: JavaBeans, COM, CORBA, RMI, EJB and the CORBA component model[C]//Proc of the 24th international conference on software engineering. [s. l.]:[s. n.],2002.

[8] Mili A, Mili R, Mittermeir R. Storing and retrieving software components: a refinement based system[C]//Proc of 16th ICSE. [s. l.]:IEEE Computer Society Press,1994:91-100.

[9] 王希辰. 可复用软件构件的表示和检索[J]. 计算机工程,2002,28(12):80-82.

[10] 付青华,林　宁,冯　惠,等. 基于刻面分类的构件检索系统的设计与实现[J]. 计算机应用与软件,2010,27(6):57-59.

[11] 任姚鹏,陈立潮,张英俊,等. 基于潜在语义分析的构件聚类改进方法[J]. 计算机工程,2011,37(4):67-69.

[12] Xie Binhong, Ren Yaopeng, Zhang Yingjun, et al. Research on the clustering algorithm of component based on the grade strategy[C]//Proc of international conference on computer application and system modeling. [s. l.]:[s. n.],2010.

[13] 田晓珍,任姚鹏,王春红. 一种改进的构件聚类索引树研究的研究[J]. 现代计算机,2014(23):12-15.

[14] 张英俊,任姚鹏,陈立潮,等. 基于语义相似度与优化的构件聚类算法[J]. 计算机工程与设计,2010,31(11):2531-2535.

(上接第 109 页)

[2] ANSI. Fibre Channel Physical and Signaling Interface (FC-PH), X3[M]. US: ANSI,1994.

[3] ANSI. Fibre Channel Avionics Environment-Anonymous Subscriber Messaging (FC-AE-ASM), Rev1.2[M]. US: ANSI, 2006.

[4] 屠晓杰,熊华钢,徐　鼎,等. 支持多种 FC-4 层协议的光纤通道接口卡实现[J]. 电讯技术,2013,53(2):188-194.

[5] 章宇东. SOC 技术在 FC 芯片设计中的应用[J]. 航空电子技术,2005,36(1):42-48.

[6] 杨海波,田　泽,蔡叶芳,等. FC IP 软核的仿真与验证[J]. 计算机技术与发展,2009,19(9):168-172.

[7] T11/Project 1237-D/1.71: Information Technology Fibre Channel-Audio Video (FC-AV)[S]. US: American National Standards Institute,2001.

[8] 王红春. 基于 FC 的航电数字视频传输技术研究[J]. 计算机技术与发展,2010,20(5):250-252.

[9] 刘　浩,田　泽. FC-AV 协议及实现方法研究[J]. 计算机技术与发展,2012,22(7):1-4.

[10] 黄浩益,黄栋杉,徐晓飞. 光纤通道技术在航电系统中的应用[J]. 航空电子技术,2005,36(3):9-14.

[11] 申　敏,曹聪玲. 基于 SoC 设计的软硬件协同验证技术研究[J]. 电子测试,2009(3):9-12.

[12] 陈孟杰,于海勋. 光纤通道 8B/10B 编解码模块设计[J]. 电子测量技术,2007,30(5):161-164.

[13] 田　靖,田　泽. AFDX-ES SoC 虚拟仿真平台的构建与应用[J]. 计算机技术与发展,2010,20(8):192-194.

[14] 郭　亮,李　玲,田　泽,等. ARINC659 总线接口芯片的 FPGA 原型验证[J]. 计算机技术与发展,2009,19(12):240-242.

[15] 孙玉焕. 64 位 CPU 的 FPGA 原型验证[J]. 现代电子技术,2007,30(21):158-160.