基于检索条件转换算法的多构件库检索①

郑立垠　郎颖莹　安　璐　(中国石油大学(华东) 计算机与通信工程学院 山东 东营 257061)

摘　要： 不同构件库之间实现互通可以有效提高复用者检索构件的效率，分类是检索的基础。通过建立多个以关键字和本体分类方式的构件库的检索条件转换模式帮助用户从基于这两种分类模式的多个构件库中检索构件，避免用户多次为同一需求构造不同的检索条件，减少复用者的理解成本，提高关键字检索本体构件库的查全率。实验结果证明了该方法的有效性和可行性。

关键词： 多构件库 构件检索 关键字分类 本体分类 条件转换

Component Retrieving in Multi-Library Based on Retrieval Conditions Transformation

ZHENG Li-Yin, LANG Ying-Ying, AN Lu

(Institute of Computer Science and Communication Engineering,China University of Petroleum, Dongying 257061,China)

Abstract: To implement sharing resources in different component libraries can improve the retrieval efficiency. Classification is the basis of retrieval. This paper sets up a number of component libraries based on keywords and ontology classification and introduces the transformation of retrieval conditions in order to help users to retrieve components from the component libraries based on the two classification models. It can avoid users' proposition of different retrieval conditions for the same demand repeatedly. It can also decrease the cost of understanding in retrieval and improve the recall for keyword retrieve ontology component library. The validity and feasibility of the method is vertified by the experiment.

Keywords: multi-library;component retrieval; keyword classification; ontology classification; conditions transformation

构件的表示与检索技术是可复用软件构件库的两个主要核心技术。随着构件库的发展，出现了多种分类检索方式，其中，REBOOT、NATO 都提出了各自的可复用软件构件的分类方案，青鸟构件库[1]所采用的也是以刻面分类为主、多种分类模式相结合的方法对构件进行分类描述。在刻面基础上表示的基于本体的构件分类检索技术也得到了广泛的重视和应用。近几年来单个构件库技术的研究和实践取得了相当成果，但是随着构件数量的加剧增多，在单一构件库查询检索构件已经不能满足复用者的需求，为了获得所需构件不得不逐一访问每个构件库，这样会导致检索效率低[2]，所以在面向不同分类方式的构件库时，存在对多构件库检索技术研究的不足。

1　引言

多个不同的构件库之间实现互通可以有效扩大重用者检索构件的范围和提高检索效率，而检索条件转换是多构件库检索亟待解决的问题。文献[3]提出了多种检索方法的检索条件转换机制，包括关键字与刻面检索条件的相互转换以及刻面条件转换到属性-值、枚举条件的检索算法。本文在此基础上，给出一种基于关键字和本体分类模式的多个构件库检索条件转换算法，该算法能实现在用户输入关键字的检索条件时，

① 收稿时间:2009-04-23

能够对本体分类模式的构件库检索构件，使用户在不清楚构件库分类模式的前提下能同时检索这两种不同分类模式的构件库。实验证明该算法为实现跨构件库检索奠定了基础，减少了重用者的理解成本，提高了检索的查全率。

2 多构件库检索

在构件检索过程中，检索对象如果是多个异质构件库，各个构件库对软构件分类检索方法可能不一样，用户如果选择一种检索方法向系统提交检索请求，这种检索方法可能只能被某一个或某一些构件库所识别，要使用户的检索信息能够被要检索的每个构件库所能理解和接受，就必须把用户的检索方法和检索条件转换成能够被各个构件库所理解的检索方法和检索条件。经过这种转换后，检索信息才能被各个构件库所理解和接受，并进一步在各个构件库中进行检索。

2.1 概述

分类模式(Classification Mode)是构件库中构件所拥有的一组共同分类特征的集合，不同构件库适应不同领域特性。构件的分类方法及相应的库结构对构件的检索和理解有着极为深刻的影响。W.Frakes 将现有的构件分类检索方法分为信息科学方法、超文本方法和人工智能方法三类。本文将以典型的关键字和本体分类检索方法为例，分析检索条件转换算法。

2.2 关键字分类

关键字分类法是一种最简单的构件库组织方法，其基本思想是：每个构件用一组能描述其基本特征的关键字编目，用户在对以关键字分类模式的构件库进行检索时通常只会输入一些与目标构件相关的一系列词语。这些词语属于不受控词语，由于缺乏上下文语境导致在检索过程中检索的效率和精确度得不到保证。

2.3 本体分类

本体分类是在刻面分类描述的基础上引入的关于某些领域的共享理解，而刻面分类是指由一组描述构件本质特征的刻面所组成，每个刻面从不同的视角对构件库中的构件进行精确的分类，每个刻面具有一组术语(关键字)，术语之间有类层次关系而形成结构化的术语空间。在这个基础上引入本体后，作用在于支持逻辑推理以及某一领域知识的共享、复用[4]。基于本体分类方式的构件库一共包括两部分，一部分是刻面子本体，构件描述仍然以刻面方案为主，另一部分是领域子本体，这一部分将作为构件描述和检索的知识基础存在。领域子本体中的知识将有助于揭示构件复用需求以及构件描述的真实含义，提高构件复用的机会。这样用户在检索的过程中通过领域子本体可以发现用户检索的潜在的含义，以及可以帮助用户发现用户想不到的方面，从这方面来讲，通过本体检索提高了构件检索的查全率。

3 多构件库检索过程模型

多构件库检索过程模型包括以下几步：

(1) 复用者分析实际的需求；

(2) 根据复用者的知识水平形成并提出查询条件；

(3) 由于各个构件库分类方法各异，需要将检索条件转换成构件库能识别的检索条件；

(4) 然后将检索条件提交到各个构件库中进行检索；

(5) 在库中通过匹配算法进行匹配，找出精确匹配或近似匹配的构件；

(6) 最后返回检索结果集。

整个检索过程如图 1 所示。

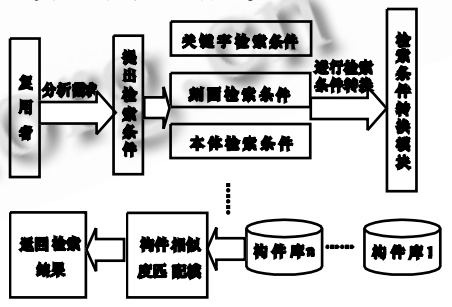

图 1 多构件库检索过程模型

本文所设计的多库检索模型采用关键字检索，不管构件库是采用关键字分类模式还是本体分类模式，都能理解这种检索条件。关键字检索适用于大部分用户，用户在输入检索条件时只需输入用户所熟悉的与要检索的构件相关的一个或多个关键字。关键字检索条件针对以关键字为分类模式的构件库来说是可以理解的，而对于以本体为分类模式的构件库来说则需要先进行检索条件转换算法，把用户在检索时输入的关键字转换成检索本体构件库时所能识别的检索条件。

具体检索过程如图 2 所示。

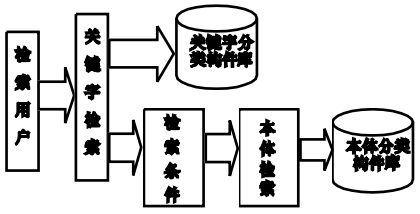

图 2　跨关键字、本体构件库检索过程示意图

4　检索条件转换算法实现

4.1 算法思想

算法总体包括五个步骤：

(1) 对用户输入的关键字集合在刻面子本体中进行匹配；

(2) 查找同义词典，对该关键字集合的同义词集合在刻面子本体里进行匹配；

(3) 根据本体构件库的特点，把前两步匹配得到的关键字集合与领域子本体中的概念进行相似度匹配；

(4) 把步骤(3)得到的概念再与刻面子本体中的刻面及术语进行匹配；

(5) 把最终得到的术语集合映射为本体所识别的 OWL 文件。

4.2 算法描述

用 K 表示用户输入的关键字集合，F 表示构件库刻面子本体的刻面集合，F(S)表示刻面下的术语集合 S，SS 表示领域子本体下的概念术语集合，Results 表示检索条件关键字集合，T 为临时存储变量集合。算法具体实现如下：

输入：一组关键字集合 $K=\{K_1,K_2,\cdots,K_k|k$ 为关键字术语个数$\}$；

输出：本体库检索条件集合；

(1) Results=K；T= Ø；

(2) for　刻面子本体中的每个刻面 Fi

(3) 判断该关键字集合 K 与 F_i 刻面下术语的映射关系 R

(4) if 此刻面下存在某术语集合 $F_i(S)$与用户输入的关键字集合 K 存在映射关系 $R\{K,F_i(S)\}$ then

(5) Results = Results$\cup\{F_i(S)\}$

(6) end for；

(7) if　Results= ?　then

(8) 查找同义词词库，查找出关键字集合 K 的同义词集合 $K'=\{K_1',K_2',\cdots,K_k'|k$ 为 K 集合同义词个数$\}$，执行(2)-(6)；

(9) for　Results 中的每个术语 Resultsi

(10) for　领域子本体下的每个概念术语 SSi

(11) 判断 Results 下的术语 Resultsi 与构件库中领域子本体下的术语 SSi 的相似度 Sim(Resultsi, SSi)

(12) if　领域子本体下存在某概念术语 SSi 与 Results 中的某一术语符合条件 Sim<t (阈值)　then

(13) T= T$\cup${SSi}；

(14) end for；

(15) end for；

(16) 对集合 T 执行(2)—(8)；

(17) 在概念词典[5]的帮助下，将 Results 中的单词映射为本体中的概念，生成检索树的 OWL 文件，即本体库检索条件集合；

(18) end

在这个算法中，关系 R 表示关键字术语集合 K 和刻面术语集合 F 之间的映射关系定义如下：若对于关键字术语集合 K，在某一刻面术语集 Fi 中存在子集 S，符合关系 R{K,Fi (S)}，则称关键字术语集可以转换为该刻面 Fi 中的刻面术语集 S。

这里的 Sim 采用文献[6]提出的语义相似度计算方法，设 t1 和 t2 是本体中的两个概念，Sim(t1,t2)表示这两个概念之间的相似程度，则有,其中 n 是术语 t1 和 t2 在本体中 kind-of 关系的层次中所具有的最大深度；是权重(可简单的取)；取值定义如下：；Sim 返回[0,1]间的实数表示两个概念间的语义相似度。其中阈值 T 从专家或经验数据获取得到，在检索系统中可以由以往的数据经验得到。

在使用了检索条件转换算法后，能极大的提高用户检索构件的查全率。例如用户在检索条件中输入“外贸付款”。表示用户需要检索有外贸业务付款功能的构件。由于在外贸电子商务领域子本体中，“信用证 L/C(Letter of Credit)”，“电汇 T/T(Telegraphic Transfer)”，“付款交单 D/P(Document against Payment)”是“外贸付款”的几种方式。所以系统在检索有外贸付款功能的构件时，首先在刻面子本体中进行“外贸付款”这个关键字的匹配，查找到相关的刻面术语后，然后查找同义词典，如果有同义词则到刻面子本体继续匹配，如果没有同义词则转到领域子

本体，通过相似度计算得知“外贸付款”和“信用证 L/C”、“电汇 T/T”及“付款交单 D/P”的语义相似度很大，于是系统在检索具有“外贸付款”功能的构件时，具有“信用证 L/C”、“电汇 T/T”及“付款交单 D/P”功能的构件也将提供给用户选择。所以说进行检索条件转换算法后能根据构件描述中没有但隐含的语义进行检索，提高了查全率。如果不使用条件转换算法，用户在单纯输入“外贸付款”这个词，仅利用关键字检索，将不能检索出具有“信用证 L/C”、“电汇 T/T”及“付款交单 D/P”功能的构件。

4.3 算法分析

通过对算法的描述，可以看到该算法具有如下特点：

(1) 该算法适用于大规模以关键字和本体这两种分类模式的构件库的构件检索，构件库数目越多，越能显示出本算法比用户逐一检索构件库的优越性；

(2) 该检索条件转换算法能有效的实现关键字检索条件到本体检索条件的转换，转换的风险程度小，可行性高，是可以采用的转换方案。

5 实验

为验证检索条件转换算法的有效性，我们实现了一个跨构件库检索原型系统，采用了检索条件转换模型，实现了由关键字到本体的检索条件转换算法。我们通过两种检索方法对现有的两个构件库进行了测试，实验设计如下：在构件库 A 中存储了 500 个以关键字描述和分类的构件的描述信息，在构件库 B 中存储了 500 个以本体描述和分类的构件的描述信息，然后模拟用户对构件进行检索，用户包括计算机专业和非计算机专业人士。在没有使用相关检索条件转换算法情况下检索，用户需要依次访问这两个构件库。共检索 100 次，每 10 次统计一次，得到的查全率大约在 0.6 左右。采用相关检索条件转换算法后，每 10 次统计一次，得到的查全率大约在 0.85 左右。图 3 是使用和不使用检索条件转换算法的查全率比较图。以上测试验证了本检索方法的可行性和有效性。

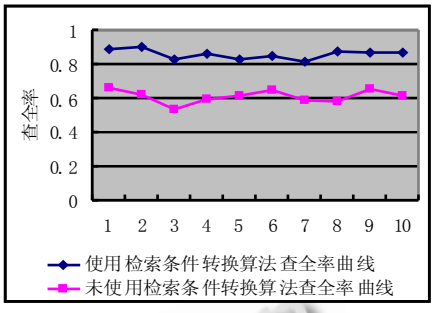

图 3 使用检索条件转换算法前后查全率比较图

6 总结与展望

本文针对关键字和本体两种不同的描述分类方式的多个构件库的构件检索进行研究，借鉴关键字和刻面检索条件相互转换的思想，结合构件库本体描述的具体特征，提出了一种关键字转换为本体检索条件的算法，给出了具体步骤，该算法可以作为跨库检索时检索条件预处理的有效方法之一，在一定程度上解决了用户跨构件库检索时检索条件不一致的问题，符合多构件库检索的思想，提高了构件检索的查全率。但是此检索条件转换算法会在一定程度上增加检索过程中的工作负荷，尽管如此，使用了这种检索条件转换算法比人工多次检索相比，它的时间代价相对还是少的，所以在以后的工作中我们还要继续研究以达到减少检索条件转换算法的工作负荷，尽可能改进算法性能。另外为了更加有效的使用和推广该算法，还需要提供友好的查询构件界面和高效的构件库平台，这都是我们下一步工作的方向。

参考文献

1 Chang JC, Li KQ, Guo LF, et al. Representing and retrieving reusable software components in JB(Jadebird) System. Electronica Journa1, 2000,28(8):20－24.

2 马亮,谢冰,杨芙清.多构件库统一刻面检索机制.电子学报, 2002,30(12A):251－254.

3 盛义芳,张维石,张秀国,史金余.面向多构件库的构件检索条件转换机制研究.计算机工程与应用, 2006:27－30,35.

4 Wang XH, Zhang DQ, Gu T, et al. Ontology based context modeling and reasoning using OWL. Proc. of the 2nd IEEE Annual Conference on Pervasive Computing and Communications Workshops (PERCOMW 2004). 2004.18－22.

5 李振东,费翔林.基于概念的信息检索模型研究.南京大学学报(自然科学), 2002,38(1):99－109.

6 朱礼军.万维网环境下基于领域知识的信息资源管理模式研究[博士学位论文].北京:中国农业大学, 2004.