多信息源下本体自动抽取的实现*

刘勇军, 聂规划
(武汉理工大学 信息管理与信息系统系, 武汉 430070)

摘 要: 以关系型数据库、XML文件、HTML文件、一般文档为信息源,运用不同的方法分别将多信息源映射为概念图,并按照拟定的概念逻辑结构进行统一存储,最后运用抽取算法实现本体的自动抽取。

关键词: 本体抽取; 语义网; 本体映射
中图分类号: TP311 **文献标志码**: A **文章编号**: 1001-3695(2007)11-0183-02

Implementation of automatic ontology extraction based on multiple information sources

LIU Yong-jun, NIE Gui-hua
(Dept of Information Management & System, Wuhan University of Technology, Wuhan 430070, China)

Abstract: Multiple information sources, for example, relational database, XML file, HTML file and document, were mapped as concept map respectively. Subsequently, they were executed unified storage according to the logical structure of concepts constucted beforehand. Finally, extraction algorithm was applied to automatically extracting ontology.
Key words: ontology extraction; semantic Web; ontology mapping

作为一种能在知识层提供知识表达和共享的概念体系,本体为语义网的实现奠定了语义基础。本体也可定义 Web服务间交互消息中的概念和意图,以提高 Web服务的表达能力和组合能力[1]。出于对各自问题域和具体工程的考虑,构造本体的过程各不相同,目前还没有一个标准的本体构造方法。现阶段的通常做法是在领域专家对信息进行提取和归纳的基础上,以手工方式构建本体。然而,本体的手工构建是一项工作量巨大并且异常繁杂的任务。在这种背景下,如何从已有的网页、数据库、文档等信息源中自动抽取领域本体,提高本体的创建效率和质量已成为近年来语义网和语义 Web服务研究的热点。

多信息源的概念映射

1)关系型数据库映射

关系型数据库的数据组织规范、结构明确。关系型数据库管理系统提供了数据表的关系视图导航表示法,通过关系视图导航工具可以直观地获取数据关系,同时可以通过生成 SQL脚本工具来生成关系脚本文件。利用逆向方法可获取数据库的数据逻辑结构,进而抽取出相关系统的本体概念。在数据库中,每个数据表可映射为概念图概念节点的实体类;数据表的属性可映射为概念图概念节点的属性类;参照关系可映射为概念图的连接类;数据表中的属性值域可映射为概念的类标号,有效性规则可映射为约束。

2) XML文件映射

DTD或 XML schema定义了 XML文档的结构和约束,规定了元素清单、属性、文档中的实体及其相互关系,为本体的抽取奠定了基础。DTD包括内部 DTD、外部 DTD和公用 DTD三种形式。

a)内部 DTD包含在 XML文件的序言中,其 XML文件结构为

〈?xml version = "1.0" encoding = "GB2312" standalone = "yes"?〉
〈!DOCTYPE 根元素名 [元素描述]〉
文件体 …

b)外部 DTD是以 DTD独立文件形式存在的,引用其 XML文件的结构为

〈?xml version = "1.0" encoding = "GB2312" standalone = "no"?〉
〈!DOCTYPE 根元素名 SYSTEM "外部 DTD文件的 URL"〉

c)公用 DTD是由权威机构制定的提供给特定行业或公众使用的 DTD。其引用形式为

〈!DOCTYPE 根元素 PUBLIC "DTD名称" "外部 DTD的 URL"〉

进行 DTD的概念映射时,内部 DTD需要截取序言中的定义部分;外部和公用 DTD需要首先获得指定的文件,然后将包含有成组元素的元素映射为概念图中概念节点的实体类。成组元素中的子元素如果不可映射为实体类,则映射为概念图节点的属性类。如果被包含的子元素可映射为实体类,则两实体类之间存在的相互关系可映射为概念图的连接类。元素内容模型中以正则表达式表示的规则映射为对应元素的约束。XML schema定义了简单类型(simpleType)和复杂类型(complexType)两种主要的数据类型。在进行本体映射时,简单类型的元素映射为概念图中节点的属性类;复杂类型的元素映射为概念图中概念节点的实体类。复杂类型的元素嵌套复杂类型的元素时,就建立了两实体类之间的关系;简单类型的约束对应着相应元素的约束。

收稿日期: 2006-10-09; **修返日期**: 2006-12-28 **基金项目**: 国家自然科学基金资助项目(70572079)
作者简介:刘勇军(1975-),男,湖北黄梅人,讲师,博士研究生,主要研究方向为知识管理、供应链协同(liuyongjun_163@163.com);聂规划(1958-),男,教授,博导,主要研究方向为知识管理、人工智能.

© 1994-2007 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3) HTML文件映射

Web是供应链内部发布和获取信息的主要渠道。WWW已成为快速增长的巨大信息库。HTML页面中部分有规律的信息可作为本体抽取的信息源。其中,页面中数据表部分可从〈Table〉〈/Table〉中抽取本体的实体类;显示表结构的〈tr〉〈/tr〉中系列的〈td〉〈/td〉映射为属性类;表中嵌套的表可映射为两实体类之间的关系。针对页面纵栏式数据部分,文本类对象的 id值可映射为概念图中节点的属性类;紧接系列属性类的对象的 id值映射为概念图中概念节点的实体类;属性类值的超链接映射为类间关系[2~4]。

4)一般文档映射

日常管理中产生了大量文档,但规律性不强,经整理后可作为本体抽取的信息源。抽取时,首先对文档进行预处理,将停用词、标点符号、英文字母、数学运算符等其他非汉字字符用空格代替;取出所有空格分割的字符串,并统计这些字符串的出现频数,去掉频数小于一定阈值的字符串,得到统计模式的词条集合候选集。采用 TFIDF(term frequency inverse document frequency)法计算词在文档中的权重,构建词—文档矩阵。权重 TFIDF的公式为

$$w_{ij} = [\, tf_{ij} \log_2 (M/m_i + 0.01)\,] \Big/ \sqrt{[\sum_{i=1}^{m} tf_{ij} \log_2 (M/m_i + 0.01)\,]^2}$$

其中:tf_{ij}表示词 i在文档 j中的频率;M表示文档集合总数;m_i表示在文档集合中出现词 i的文档数目。利用奇异值分解法可将由 w_{ij}构成的矩阵 W分解为 $W = U \times S \times V^T$。其中:U和 V分别为 $m \times m$和 $n \times n$的正交矩阵;S为对角矩阵;S的非零对角元素 δ_i ($i = 1, \cdots, r$)为矩阵 W的奇异值;r为非零对角元素的个数。用 W的 k秩近似矩阵 W_k替代 W,将奇异值按从大到小顺序排列,保留矩阵 S中奇异值的前 k个,保留矩阵 U、V中与前 k个奇异值相对应的前 k列,可以得到 $W_k = U_k \times S_k \times V_k^T$。其中:$U_k$代表了词和概念之间的关联,可将每一列映射成一个概念,列的各项元素暗含了词之间的关联;V_k代表了概念和文档之间的关联[5]。

概念图存储

在获取概念关系的前提下,可以将多信息源映射为概念图,并按照拟定的概念逻辑结构统一存储以便供抽取本体之用[6]。在进行存储时所依据的原则有:

a)一个概念节点的逻辑结构可以用一个记录表示。这个记录有概念标记 ID、概念标记、概念类型三个属性。概念标记 ID是概念系统中所给出的惟一编码;概念标记是指概念的规范名称;概念类型是概念的类别说明,包括实体类、属性类、实体子类、泛化类等。

b)属性类的有效性规则可用具有属性 ID、概念标记 ID、属性名、所指域和规则表达式五个属性的记录表示。属性 ID给定属性的惟一标志号;概念标记 ID依从于概念节点;属性名为概念属性的规范命名;所指域是该属性必须有的值的集合;规则表达式则是使用正则表达式语法定义的规则模式。

c)一个关系节点可用具有关系 ID、关系名和关系类型三个属性的记录表示。关系 ID是关系的惟一标志码;关系名对应于概念之间联系的规范名称;关系类型表明关系的类别,包括 partof、kindof、instanceof、attributeof等关系。

d)一个连接概念节点和关系节点的记录具有概念标记 ID、关系 ID和关系标记三个属性。概念标记 ID依从于概念节点,关系 ID依从于关系节点。

通过连接概念节点和关系节点的记录可建立概念节点与关系节点的连接关系。这样的结构符合面向对象的关系数据库设计准则,是相对稳定的,可以满足数据搜索的需求。概念图逻辑存储结构视图如图 1所示。

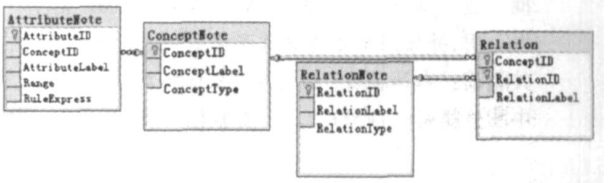

图 1 概念逻辑结构存储视图

本体自动抽取算法

概念逻辑结构以数据表方式存储,概念之间关系映射为数据表之间的联系。这样在抽取本体时就可利用 SQL语句直接进行运算。首先定义 $A(c)$为概念 c的所有属性的集合;res(a)表示属性类 a的约束;range(a)表示属性类 a的值域;$R(c, a)$表示概念 c具有属性 a,或者 a是概念 c的属性;$C(r)$表示关系 r所联系的概念类的集合;$R(C(r))$表示 $C(r)$中的概念间关系为 r[7]。本体抽取算法的具体步骤如下:

a)提取概念进入概念集合 $C(c)$

//select ConceptLabel from ConceptNote;

b) for each Concept c in $C(c)$ do

提取属性进入概念属性集合 $A(c)$ /* select AttributeLabel from AttributeNote, ConceptNote where AttributeNote.ConceptID = ConceptNote.ConceptID and ConceptLabel = c */

提取规则进入约束集合 res(a) /* select RuleExpress from AttributeNote, ConceptNote where AttributeNote.ConceptID = ConceptNote.ConceptID and ConceptLabel = c */

提取属性的值域进入值域集合 range(a) /* select Range from AttributeNote, ConceptNote where AttributeNote.ConceptID = ConceptNote.ConceptID and ConceptLabel = c */

提取属性关系进入关系集合 $R(c, a)$

调用泛化函数 Find_Abstract_Concept($A(c)$)得到泛化类、类间关系、转移约束

调用分类函数 Find_Sub_Concept($A(c)$)得到子类、类间关系;

c)提取概念间关系进入类关系集合 $R(r)$

//select RelationLabel from RelationNote;

d) for each Relation r in $R(r)$ do

提取关系所联系的概念集 $C(r)$ /* select ConceptNote.ConceptLabel from Relation, RelationNote where RelationNote.RelationLabel = r and Relation.RelationID = RelationNote.RelationID and Relation.ConceptID = ConceptNote.ConceptID */

构造概念关系 $R(C(r))$

提取子关系和互反关系 /* select Relation.RelationLabel from RelationNote, Relation where RelationNote.RelationLabel = r and Relation.RelationID = RelationNote.RelationID */;

f)选用本体描述语言表示类、关系、约束、值域。

其中,泛化函数 Find_Abstract_Concept($A(c)$)通过放宽属性的约束和值域而得到更加抽象的父类。其计算过程如下:

(下转第 187页)

© 1994-2007 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

应用本算法获得初步排序结果如图 1所示。图中编号 i-j 表示工件 i的第 j道工序,即工序 P_{ij};总加工时间为 75。

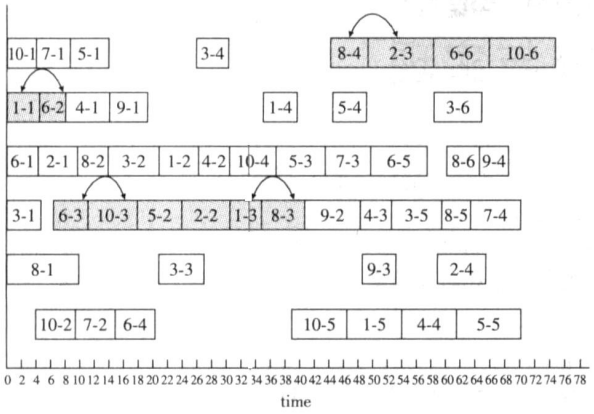

图 1　关键路径、关键块和可能的工序交换

应用优化排序得到的最优结果如图 2所示。总加工时间为 67。

结束语

本文分析了模具车间的生产特点,提出了一种基于最短加工时间和最小等待时间等组合规则的调度算法。本文实现了当工件的每道工序可在一台或多台设备上进行加工时,如何挑选机器的问题;同时考虑了不同工序的加工冲突和不同设备的资源分配问题。本算法在传统的 job shop调度问题的基础上,放宽了资源约束条件,更符合大多数企业的实际生产情况,具有实际使用价值。调度实例表明,该方法能有效满足模具车间生产的实际需要。

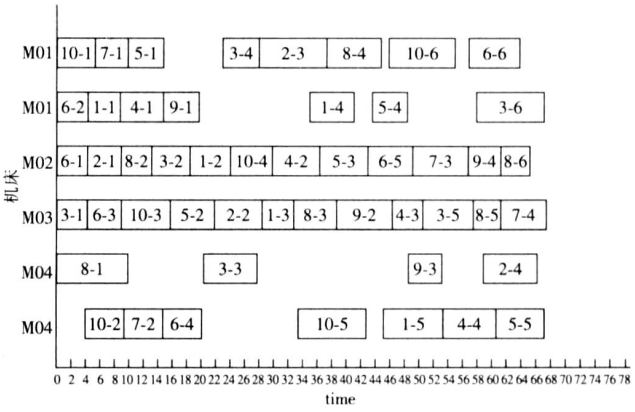

图 2　应用优化排序得到的最优结果

参考文献:

[1] 玄光男,程润伟. 遗传算法与工程优化 [M]. 北京:清华大学出版社, 2004.

[2] 张良扬. 单件小批生产模式下的计划系统与优化算法研究 [D]. 南京:南京航空航天大学, 2004: 80-100.

[3] 许文砚. 基于单件小批生产模式的车间调度方法研究和系统开发 [D]. 南京:南京航空航天大学, 2005: 40-100.

[4] 王凌. 混合优化策略和神经网络中若干问题的研究 [D]. 北京:清华大学, 1999: 50-100.

[5] 潘全科. 一类解决作业车间调度问题的遗传退火算法 job-shop调度问题 [J]. 自然科学报, 2005, 20 (5): 10-20.

[6] 何霆. 车间生产调度问题研究 [J]. 机械工程学报, 2000, 36 (5): 97-103.

(上接第 184页)

当未定义初始属性 x时,令 $x = a$

for each Attribute x in $A(c)$ do

获取属性 x的约束 res(x)和值域 range(x),放宽属性的约束和值域得到新的抽象类 c以及该类下的属性 x的新约束 res(x)和新值域 rang(x)

抽取子父类关系 $R(c, c')$

以 x为初始属性,调用泛化函数 Find_Abstract_Concept($A(c')$)

从 $A(c)$中除去属性 x的约束 res(x)和值域 range(x)

相反,分类函数 Find_Sub_Concept($A(c)$)则是通过收缩属性的约束和值域来得到更加具体的子类。其计算过程如下:

当未定义初始属性 x时,令 $x = a$

for each Attribute a in $A(c)$ do

获取属性 x的约束 res(x)和值域 range(x),收缩属性的约束和值域得到新的抽象类 c以及该类下的属性 x的新约束 res(x)和新值域 range(x)

抽取父子类关系 $R(c, c')$

以 x为初始属性,调用分类函数 Find_Sub_Concept($A(c')$)

从 $A(c)$中恢复属性 x的约束 res(x)和值域 range(x)

$A(c) = A(c) - c$

结束语

本文在分析关系型数据库、XML文件、HTML文件和一般文档特征的基础上,提出了多信息源下本体的自动抽取方法。这种方法有利于克服单一信息源下本体抽取的不完整性以及发挥各类信息源在本体抽取中的优势,提高本体的创建效率和质量。同时,本体的自动抽取过程也需要人工进行检验和完善,以避免抽取的本体中存在着重复和矛盾。此外,本体自动抽取的学习机制、本体的一致性检测、本体映射等问题还有待于进一步研究。

参考文献:

[1] KERRIGAN M. WSMOViz: an ontology visualization approach for WSMO [C]//Proc of the Information Visualization. Baltimore: IEEE Computer Society, 2006: 411-416.

[2] 成瑜,何洁月. 本体驱动的半结构化 Web生物数据抽取 [J]. 计算机工程, 2006, 32 (5): 192-194.

[3] 王放,顾宁,吴国文. 基于本体的 Web表格信息抽取 [J]. 小型微型计算机系统, 2003, 24 (12): 2142-2146.

[4] TANAKA M, ISHIDA T. Ontology extraction from tables on the Web [C]//Proc of the 2005 Symposium on Applications and the Internet. Washington D C: IEEE Computer Society, 2006: 284-290.

[5] 董慧,余传明. 中文本体的自动获取与评估算法分析 [J]. 情报理论与实践, 2005, 28 (4): 415-418.

[6] 黄伟,金远平. 形式概念分析在本体构建中的应用 [J]. 微机发展, 2005, 15 (2): 28-31.

[7] 马峻. 一种从线性概念图中自动抽取本体概念的算法 [J]. 计算机工程与应用, 2004, 40 (23): 161-164.

© 1994-2007 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net