◎研究、探讨◎

实例驱动的自适应本体学习

张瑞玲,王文斌,王秀峰,陈秋双
ZHANG Rui-ling,WANG Wen-bin,WANG Xiu-feng,CHEN Qiu-shuang

南开大学 信息技术科学学院,天津 300071
College of Information Technical Science,Nankai University,Tianjin 300071,China
E-mail:zhangruiling@mail.nankai.edu.cn

ZHANG Rui-ling,WANG Wen-bin,WANG Xiu-feng,et al.Instances driven adaptive ontology learning.Computer Engineering and Applications,2009,45(28):31-34.

Abstract: To solve problems in knowledge management,an adaptive ontology learning approach based on metadata of knowledge resources is proposed by integrating clustering algorithm and ODP (Open Directory Project).Ontology concepts are generated by clustering documents based on their metadata,and concept hierarchy is formed based on the hierarchy of mapped concepts in ODP.In order to track the changes in knowledge,adaptive ontology learning is conducted to update the ontology and enrich ontology concepts based on the changes of cohesion and correlation of clusters.The adaptive ontology learning approach proposed in this paper can reflect evolution process and tendency in research area,and meet demands for knowledge management of knowledge organization and knowledge sharing of researchers.Experimental result demonstrates the validity of the approach.
Key words: knowledge management;ontology learning;Open Directory Project(ODP);clustering

摘 要:针对知识管理中本体构建存在的问题,将聚类算法与ODP(Open Directory Project)目录有机结合,给出了一种基于知识资源元数据的自适应本体学习方法。根据元数据对文档进行聚类形成本体概念,将生成的概念分别映射到ODP中确定概念间的层次关系,生成初始本体;根据内聚性和相关性的变化进行自适应本体学习,实现本体更新和概念丰富,以及时跟踪知识的变化。提出的自适应本体学习方法能够很好地反映研究领域的演变过程和发展趋势,满足知识型组织进行知识管理和研究人员共享知识的需求。实验结果表明了方法的有效性。
关键词:知识管理;本体学习;开放式目录项目;聚类
DOI:10.3778/j.issn.1002-8331.2009.28.009 **文章编号**:1002-8331(2009)28-0031-04 **文献标识码**:A **中图分类号**:TP181

1 前言

随着知识经济时代的来临,知识型组织将主要通过知识而不是金融资本或自然资源来获取竞争优势。相对于传统行业,知识型组织更强调基于创新的知识生产与基于共享的知识交流,因此知识管理成为知识型组织的核心要素。现有知识管理系统存在两个典型问题,一是缺少共享的知识模型,容易造成对同一知识描述不同,影响用户的理解和知识共享;二是缺少统一的知识存储形式,妨碍人们对知识的认识和交流,形成知识孤岛[1]。

本体作为一种语义和知识层次上的概念模型,具有知识共享、重用等优点,将其引入知识管理系统能很好地解决上述问题。本体通常由领域专家构建,常用的本体构建方法有TOVE法、骨架法、IDEF-5法、循环获取法、七步法等[2]。虽然目前本体构建工具已经较为成熟,但本体构建仍是一项繁琐而辛苦的任务,半自动、自动的本体生成方法成为研究的热点。Khan等从字典中抽取感兴趣的概念和关系,构建需要的本体[3],通过该方法建构的本体通常为一般性的描述,并不是与特定领域相关的本体。文献[4]提出半自动的本体构建方法,利用概念语义矩阵从文档中抽取概念并由领域专家确定概念在本体中的位置。但该方法采用UNL(Universal Networking Language)描述本体模型,而不是OWL,致使该应用不具备广泛性。Thanh等人提出模糊聚类和形式概念分析相结合的模糊本体自动生成方法,但通常生成的形式概念众多,形式网格相当复杂[5]。

对知识型组织来说,本体中的组织、人员等概念及其属性相对固定且容易确定,而研究领域信息是不容易直接获取的,而且不同类型的知识型组织所管理的知识结构千差万别,相同类型的不同组织所管理的内容也不尽相同。因此,为了更好地进行知识管理,需要根据知识型组织所拥有的知识资源建立适

基金项目:天津市科技发展计划资助项目(No.06YFGZGX05900)。
作者简介:张瑞玲(1985-),女,硕士研究生,主要研究方向:机器学习,语义Web;王文斌(1982-),男,硕士研究生,主要研究方向:语义Web;王秀峰(1941-),男,教授,主要研究方向:计算智能、金融信息系统建模与预测;陈秋双(1966-),女,教授,博士生导师,主要研究方向:供应链管理与物流系统优化。
收稿日期:2008-11-11 **修回日期**:2009-01-22

合自身需要的个性化本体,即与知识相适应的研究领域概念层次。对此,目前常用的本体构建方法是很难做到的。另一方面,知识管理系统不是静止不变的,随着知识的不断更新,如何通过学习使本体适应知识的变化并通过知识变化及时正确地判断知识发展趋势,也是亟待解决的问题。

针对知识管理中本体构建存在的上述问题,给出了一种根据知识元数据建立研究领域概念层次,以及基于知识增量的自适应本体学习方法。

2 本体及概念层次简介

2.1 本体

本体可以表示成一个五元组的集合:$O=(C,R,\sigma,A,I)$,其中 C 为概念集合,也称为类的集合;R 为概念之间关系的集合;σ 为函数,实质上是一种特殊关系,即通过其他关系可以唯一获得的关系;A 为本体中的公理集,用来说明函数之间或关系之间存在的关联和约束;I 表示本体中实例集合[6]。一般人们最关心的是本体中的概念、关系和实例。在概念关系中,is-a 关系又是最为重要的,即两概念是子概念与父概念的关系。在下文中,仅考虑概念、is-a 关系和实例,将本体表示简化为 $O=(C,is-a,I)$。

2.2 概念层次

概念层次是以层次的形式和偏序的关系组织的概念集合,通常用层次树来表示一个概念层次,即概念层次树,树的结点表示概念,树枝表示偏序[7]。仅考虑概念间的从属关系时,树上的子结点与父结点的关系即为"is-a"关系。一个概念层次树也是一个本体。

ODP(Open Directory Project)即开放式分类目录搜索系统,是目前网上最大的人工编制的分类检索系统,是由美国加州的一名程序员 Rich Skrenta1998 年 6 月创建的(http://www.dmoz.org/)。目前 ODP 由来自全球的 72 210 名志愿编辑者维护管理,共设有 59 万多个类目,收录了 529 多万个站点,尤其在一些边缘学科或冷门学科上,其类目数量要比 Yahoo!提供的全面得多。由于 ODP 的全面、权威,目前 Google、Netscape、Dogpile、Thunderstone、Linux 等搜索引擎都在使用 ODP 的目录体系。ODP 中的类目是具有层次结构的,将类目看作概念,ODP 就是一个概念层次。

3 研究领域概念层次的生成

知识管理系统的知识资源绝大部分是以文档格式存在的,而采用文本挖掘等技术可以抽取出文档元数据。根据知识的元数据可以对知识进行聚类,聚类生成的每一个簇作为一个概念,用相应簇的中心来表示。采用 ODP 作为参考概念层次,将生成的概念分别映射到 ODP 中,根据 ODP 中被映射概念间的关系确定概念间的层次关系。

3.1 知识的特征表示形式

在本体中,每一篇文档内容都可由相应的元数据来描述,综合考虑文档的标题、摘要、关键词等元数据,将文档表示为特征向量的形式。在计算其特征权重时,根据每个元数据的重要程度赋予不同的权值,如设定标题所包含的特征的权值为 3,摘要为 2 等。

3.2 聚类技术与评价指标

所谓聚类,就是将物理或抽象对象的集合分组成为由类似的对象组成的多个类的过程。在具体应用时,聚类算法的选择取决于数据的类型和聚类的目的。除此之外,聚类结果的优劣还依赖于所用聚类算法的参数值。一般情况下,用户对所要处理的数据集缺乏先验知识,甚至一无所知,要想一次就确定出合适的参数从而得到最佳的聚类结果一般是不可能的,通常需要在不同的参数下对数据集进行重复挖掘,这就是所谓的参数改变的重复聚类方法[8]。该文在生成概念层次时采用了参数改变的 k-means 聚类方法[8],以得到理想的聚类结果。

聚类算法对数据对象进行分簇,使得同一簇内的对象尽可能相似,不同簇之间的数据对象尽可能相异。借鉴文献[8]的方法,采用"内聚性"指标衡量同一个簇内数据对象的相似程度,"相关性"指标衡量隶属于不同簇对象的相关程度。

3.2.1 内聚性

在聚类分析中通常采用簇的平均半径来描述簇内数据对象的相似性,平均半径越小,簇内对象越相似,内聚性越大。簇 C_i 的平均半径定义为:$\overline{R}_i=\frac{\sum_{p\in C_i}|p-O_i|}{n_i}$,其中 O_i 表示 C_i 的中心;n_i 表示 C_i 中对象的个数;$|p-O_i|$表示 p 到 O_i 的距离,采用余弦相似度度量距离。簇 C_i 的内聚性计算公式为:

$$CI_i=\frac{1}{\overline{R}_i}=\frac{n_i}{\sum_{p\in C_i}|p-O_i|} \tag{1}$$

3.2.2 相关性

聚类结果中簇与簇之间的相关程度主要依赖于两个簇的相邻边界,可以通过以下两个方面度量:

(1)两个簇相邻边界的距离和簇大小的比率

相对于两个簇的大小来说,它们最相邻的边界相距较远,则可以认为相关性较小,簇 C_i、C_j 相邻边界的距离和簇大小的比率 $BorderDis$ 计算如下:

$$BorderDis(i,j)=\frac{\overline{R}_i+\overline{R}_j}{Dis(O_i,O_j)} \tag{2}$$

其中 $Dis(O_i,O_j)$表示簇中心 O_i、O_j 的距离。

(2)两个簇间相关数据集合 S 的大小

设簇 C_i 中距离中心 O_i 最远的数据对象用 p_{ifar} 表示,$R_i^{\max}=|p_{ifar}-O_i|$为簇 C_i 的最大半径,用 $S=\{p|p\in C_j,|p-O_i|\leqslant R_i^{\max}\}$表示簇 C_i 与簇 C_j 的相关数据的集合。当 $S\neq\varnothing$ 时,表明 C_j 边界附近的数据对象相对于 p_{ifar} 来讲,距离 O_i 更近,表明两个簇的相关性较大。用$|S|$表示集合 S 中的数据对象个数,簇间相关数据集合对相关性的影响大小可用式(3)表示:

$$MixRatio(i,j)=\frac{\sum_{p\in S}\frac{|p-O_i|}{|S|}}{R_i^{\max}}*\frac{|S|}{n_j} \tag{3}$$

将上述指标求加权和,可得到簇 C_i、C_j 相关性的计算公式:

$$CoR(i,j)=\alpha\cdot BorderDist(i,j)+(1-\alpha)\cdot MixRatio(i,j) \tag{4}$$

其中 α 为权值,满足 $0\leqslant\alpha\leqslant1$。

综合考虑内聚性和相关性,用如下指标 E 评价具有 k 个簇

的聚类结果的质量:

$$E=\sum_{i=1}^{k}\frac{1}{CI_i}+\sum_{i=1}^{k}\sum_{j=1}^{k}CoR(i,j) \quad (5)$$

E 值越小,聚类结果越好。

3.3 聚类后生成的簇到 ODP 的概念映射方法

在聚类中,通常以簇的中心代表一个簇。簇中心可由一系列特征来描述,记为 $O_j=<f_1,f_2,\cdots,f_n>$,$f_1,f_2,\cdots,f_n$ 表示特征。特征的类型由簇对象决定,当对文本聚类时,特征一般为词语或概念。为了以下叙述方便,首先给出树的高度和树的中间结点的定义:

定义 1 树的高度($length$)等于最底层叶子结点的深度加 1。其中,树中结点 M 的深度($depth$)是指从根结点到 M 的路径长度。

定义 2 树中深度为$\lceil\frac{length}{2}\rceil$的结点称为树的中间结点,$\lceil\rceil$表示向上取整。

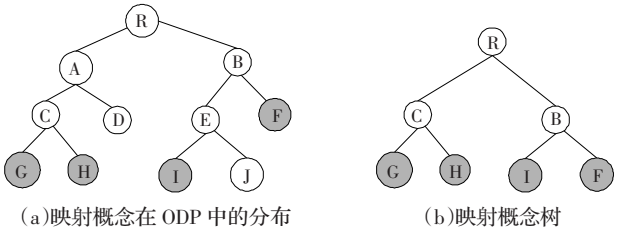
将簇 C_j 映射到 ODP 中时,首先将 n 个特征 $f_1,f_2,\cdots,f_n$ 按照字符串匹配的方法分别映射到 ODP 中的 $m(m\leqslant n)$个概念 $hc_1,hc_2,\cdots,hc_m$;然后根据 m 个概念在 ODP 中的层次关系构建映射概念树(Mapped Concept Tree,MCT);最后,根据映射规则确定选择哪个概念作为簇的映射概念。

簇的映射概念树的构建步骤:

(1)初始化 MCT 中概念集合 $A=\{hc_1,hc_2,\cdots,hc_m\}$。

(2)在 ODP 中找出 $hc_1,hc_2,\cdots,hc_m$ 的最近公共父结点 R(即所有公共父结点中深度最大的结点),$A=A\cup\{R\}$。

(3)在 ODP 中得到以 R 为根结点的子树(到 $hc_1,hc_2,\cdots,hc_m$ 中深度最大的那一层),如图 1(a)所示,其中阴影部分为 $hc_1,hc_2,\cdots,hc_m$。自下向上对子树中不属于 A 的任意概念 c,$c\in HC$,其中 HC 是 ODP 中的概念集合,利用下述方法判断是否将其加入 MCT 中。如果概念 c 的直接子结点中有两个或两个以上属于集合 A,则将 c 加入 MCT 中,即 $A=A\cup\{c\}$;如果 c 的直接子结点属于 A 的个数小于 2,但 c 的所有子结点属于 A 的个数大于或等于 2,且这些结点任两个都不存在父子关系,则 $A=A\cup\{c\}$。



(a)映射概念在 ODP 中的分布 (b)映射概念树

图 1 映射概念树构建示例

(4)对集合 A 中的概念,根据在 ODP 中的位置确定它们的父子关系,完成概念映射树的构建,如图 1(b)所示。

设定映射概念树的高度阈值为 δ(取值在具体实验中确定,$\delta\geqslant1$),则簇到 ODP 概念的映射规则为:若映射概念树 MCT 高度小于阈值 δ,以 MCT 的根结点作为簇的映射概念,否则,以 MCT 的中间结点作为簇的映射概念。

3.4 研究领域概念层次生成算法

给定知识资源集合 $D=\{d_1,\cdots,d_N\}$,利用该节给出的领域概念层次生成算法,就可得到研究领域概念层次 $RAHC=(RA,is-a)$,RA 是研究领域概念集合。算法步骤如下:

(1)对 D 的所有对象采用参数改变的重复聚类方法进行聚类,得到 k 个簇 $Cluster=\{C_1,\cdots,C_k\}$;

(2)利用 3.3 节给出的方法,将 $C_i(1\leqslant i\leqslant k)$映射到 ODP 中,得到相应的映射概念 $mc_i\in HC$,令映射概念集合 $C_{mapped}=\{mc_i|1\leqslant i\leqslant k\}$,初始化 $RA=C_{mapped}$;

(3)对 $\forall c_p\in HC$,且 $c_p\notin C_{mapped}$,如果概念 c_p 的直接子概念中有两个或两个以上属于集合 C_{mapped},则将概念 c_p 添加到 RA,即 $RA=RA\cup\{c_p\}$;

(4)根据集合 RA 中概念在 ODP 中的层次结构确定概念在 RAHC 中的层次关系,得到所研究的学术领域概念层次 RAHC。

4 基于知识增量的自适应本体学习方法

由于知识的不断发展和更新,学术资源数量正在迅速增长,在新增一定数量的实例后,原聚类的结果可能会发生局部变化。概括起来,主要有以下几种变化情形:

(1)新簇的产生:新增加的实例离原来的簇较远,不能归于原来的任何一类(簇)。这些实例很可能意味新的研究方向的产生。

(2)簇的合并:新增加的实例大都位于原来两个或多个簇之间的位置,使原有的簇的中心发生偏移,逐渐靠近,导致原来邻近的簇合并在一起。这表明:不同学科逐渐融合,形成了明显的交叉。

(3)簇的分裂:由于增加新的实例,使得原来结构稀疏的某个簇分裂成两个或多个簇。表明:随着研究的深入,研究方向逐渐明确并进一步细化,出现了新的增长点。

(4)簇中心的飘移:簇的中心随着新实例的增加而发生较大的偏移。表明:研究兴趣发生一定改变。

根据以上分析,给出了一种自适应本体学习方法 AOLA(Adaptive Ontology Learning Approach)。AOLA 能够充分反映以上变化。对于每个新产生的类,AOLA 将记录类的两部分属性:(1)时间戳,类产生的时间;(2)产生方式,分为以下三种类型:新产生(new)、原来类合并得到(merge)与原来类分裂得到(split)。

自适应本体学习方法 AOLA 的步骤:

设定阈值 a_1、a_2、a_3、a_4、n_{update}、n_{LB}。

(1)对加入的新实例,根据距离最近原则,将其标注到距离最近(相似度最大)的类[9]。如果该实例距离所有类都较远(即最小距离大于阈值 a_1 或最大相似度小于阈值),将其加入不能归类(不能标注)的实例集合 $S_{exception}$。

(2)当新增加的实例数达到 n_{update} 时,重新计算每个簇的中心。对于簇中心漂移较大的簇(原簇中心与更新后簇中心之间的距离大于阈值 a_2 的簇)修改其类名和特征。

(3)根据更新后的簇中心,重新判断 $S_{exception}$ 中的实例是否可标注,如可以,进行标注。

(4)合并:计算簇之间的相关性。当簇与簇之间的相关性超过阈值 a_3 时,将这些簇合并为一簇。

(5)分裂:对于内聚性较差(内聚性指标低于阈值 a_4)且新增实例较多的簇,对该簇内对象进行重新聚类,实现簇的分裂。

(6)引进新类:将 $S_{exception}$ 中的实例进行聚类,如果得到的簇

中实例数超过 n_{LB} 且内聚性达到要求,产生新类,作为研究领域的子类。

5 实验及结果分析

从某高校教师、学生发表论文中选择 196 篇文档,抽取出标题、摘要、关键词、作者、发表日期、出处等元数据,并根据元数据将每篇文档表示成特征向量形式。对选定的文档进行聚类,并将聚类的簇与 ODP 中的概念建立映射,形成所研究领域概念层次。主要步骤和结果如下:

(1)采用参数变化的 k-means 重复聚类算法,将 198 篇文档聚成了 8 个簇。

(2)将聚类生成的 8 个簇与 ODP 中的概念建立映射并确定概念间的层次关系(实现中参数 δ=3)。簇的映射结果如表 1 所示,形成的概念层次结果如图 2 所示。

表 1 簇的映射结果

簇	权重最大的 3 个特征	映射到 ODP 中的概念
1	multimedia,digital,mpeg	Multimedia
2	communication,telephony,cable	Data_Communications
3	algorithm,simulation,synthesis	Algorithms
4	intelligent,fuzzy,set	Fuzzy
5	data_mining,representation,extract	Data_mining
6	web_based,web,weblog	Web
7	search,search_engine,interface	Searching
8	embed,linux,windows_2000	Operating_Systems

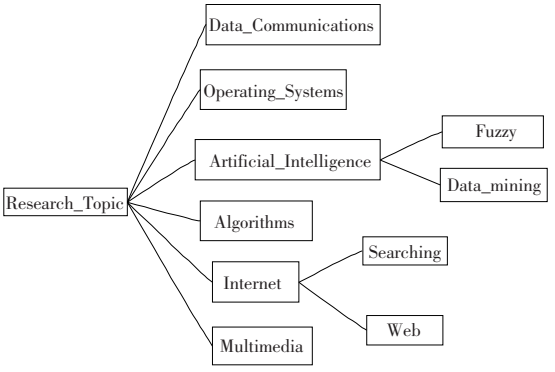


图 2 研究领域概念层次

按照发表时间,陆续添加新文档到本体实例集中,对加入新实例的本体运用 OntoIL 进行本体学习。实现过程中,设定参数 n_{update}=50,n_{LB}=10,相似度阈值 a_1=0.01。向本体实例集中添加文档 134 篇,通过标注之后,类 Internet、Searching、Web 中实例数目增加较多,且增加实例后造成 3 个簇间相关性变大达到合并条件,将属于 3 个类的实例进行重新聚类,通过参数改变的重复聚类确定最佳聚类簇数为 4, 生成的 4 个簇分别映射到 ODP 中的概念 Internet、Searching、Semantic_Web、Wireless_Sensor_Network。通过本体学习研究领域概念层次发生了改变,结果如图 3 所示。

由图 2 可以看出,根据抽取的文档元数据可以生成合理的研究领域概念层次。由图 3 可以看出,经过本体学习后,概念层次中新增加了语义网和无线传感器网络两个概念,反映了语义网和无线传感网络作为信息科学领域中新的发展方向,成为该高校新的研究方向。

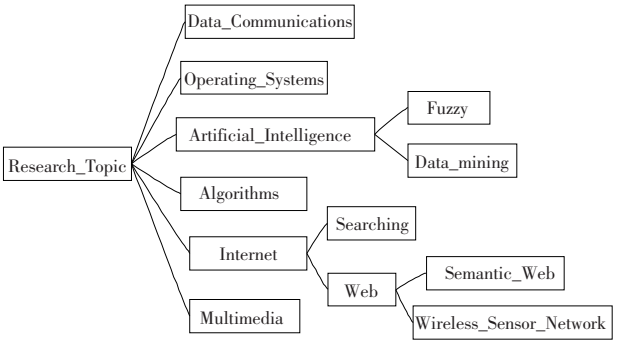


图 3 本体学习后研究领域概念层次

6 结论

将本体技术引入知识管理可以解决知识型组织缺少共享知识模型的问题,但本体的构建过程是十分复杂的,且现有本体构建技术不能解决知识管理系统中个性化本体构建的需求。将聚类技术与 ODP 目录结合, 提出了一种适应于知识资源的本体研究领域概念层次生成方法,由此构建的知识本体更符合管理者的要求,提高知识管理效率。知识型组织中的知识是不断更新的,针对这一点,给出了基于知识增量的自适应本体学习方法,实现了本体概念的丰富,使本体能够适应知识的变化,及时反映研究领域的发展趋势和最新研究动态。实验表明了方法的有效性。该方法在高等院校、科研院所等单位的知识管理系统中具有良好的应用前景。

参考文献:

[1] 徐丽平.基于本体的知识管理系统研究[J].电脑应用技术,2007(2).

[2] 李景,孟连生.构建知识本体方法体系的比较研究[J].现代图书情报技术,2004(7):17-22.

[3] Khan L,Luo Feng.Ontology construction for information selection[C]//Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence(ICTAI'02),2002:122.

[4] Zhou L P,Zhang D F,Chen X,et al.A method for semantics-based conceptual expansion of ontology[C]//Proceedings of the 15th Annual Workshop on Selected Areas in Cryptography,2008.

[5] Quan T T,Hui S C,Cao T H.Automatic fuzzy ontology generation for semantic web[J].IEEE Transactions on Knowledge and Data Engineering,2006,18(6):842-856.

[6] 宋峻峰.基于本体的信息检索模型研究[J].南京大学学报,2005,41(2):189-197.

[7] Chen Yi-fan,Xue Gui-rong,Yu Yong.Advertising keyword suggestion based on concept hierarchy[C]//Proceedings of ACM WSDM 2008,Stanford University,Stanford,California,USA,2008.

[8] 刘赏.参数改变的重复聚类问题研究[D].天津:南开大学,2006.

[9] Kushilevitz E,Ostrovsky R,Rabani Y.Efficient search for approximate nearest neighbor in high dimensional spaces[J].SIAM Journal on Computing,2000,30(2):451-474.