文章编号:1000-2367(2020)01-0024-09 DOI:10.16366/j.cnki.1000-2367.2020.01.005

一种基于关联规则与 K-means 的领域本体构建方法

李征[1,2],李斌[1]

(1.河南大学 计算机与信息工程学院,河南 开封 475004;
2.三峡大学 湖北省水电工程智能视觉监测重点实验室,湖北 宜昌 443002)

摘 要:随着网络上服务资源的规模化增长,如何帮助用户找到所需服务是一个关键问题.研究发现领域本体的构建可帮助用户有效解决该问题,而已有的一些构建方法一般依靠人工,费时费力.针对该问题,提出一种基于关联规则和 K-means 的领域本体构建方法.该方法首先利用支持向量机进行面向领域的服务分类,从分类得到的领域知识中选取初始领域概念;然后根据关联规则和 K-means 算法挖掘概念间关系,以得到初始领域本体,并使用 Wordnet 对其进行语义丰富.最后,通过 ProgrammableWeb 网站提供的真实服务集进行实验验证.实验结果表明所提出的领域本体构建方法能够识别概念间关系,进而为 Web 服务语义查询提供相应支持.

关键词:服务分类;关联规则;K-means;领域本体

中图分类号:TP391 **文献标志码**:A

随着软件即服务与面向服务的架构技术的发展,互联网上的 Web 服务资源呈现规模化增长趋势,如:ProgrammableWeb[1](简称 Pweb)网站发布的 Web API 一直在持续增长.但是,面对网络上大量、日益增长的 Web 服务,如何帮助用户准确找到其所需的服务仍然是一个关键问题.

本体是概念模型的明确的规范说明[1].目前,本体已经被广泛应用于语义 Web、智能信息检索、信息集成等领域[2].在服务信息检索过程中,根据本体知识的相关性,帮助用户进行相关性查询,进而找到更好满足用户需求的服务.

目前针对领域本体的构建方法有很多.文献[3-4]提出借助 Wordnet 或 Wikipedia 获取概念间的关系进而构建领域本体,提高了本体构建的速度与效率,但是对词典的依赖程度较高;文献[5]从 Web 数据中提取领域概念,然后利用字典(如 Wordnet)和领域概念库确定领域概念与实例,最后利用信息抽取的方法构建概念分类体系;文献[6]基于现有的领域本体,根据用户的需求动态添加、删减构建新的领域本体;文献[7]从非结构化的文档中抽取概念,然后利用语形学构建概念间的关系,但需要领域专家进行概念与关系的修正;文献[8]利用短语中词之间的关系建立概念间的关系,但是在处理本体片段映射时需要领域专家的参与.这些方法主要利用已有的组织关系构建领域本体,使用的前提是预先拥有足够丰富的领域术语以及术语间的关系,同时,需要领域专家的高度参与.基于 Web 服务描述挖掘初始领域概念并通过关联规则和 K-means 构建初始领域本体,然后借助 Wordnet 中的词汇关系对初始本体中的概念及关系进行丰富.

文献[9]提出基于构建的实体-属性矩阵利用形式概念分析生成概念格来挖掘概念层次关系,进而构建本体概念间的关系;文献[10]基于形式概念分析,在构建形式背景中利用信息熵对属性进行归约处理,提高了本体构建的质量.文献[11]通过计算找到“代表单词”,将“代表单词”作为下一次进行 LDA(Latent Dirichlet Allocation)的输入,自底向上层层构建得到初始本体,然后基于参数组合模式的规则定义,对本体

收稿日期:2018-06-15;**修回日期**:2018-12-29.

基金项目:国家重点基础研究发展计划(973)(2014CB340404);国家自然科学基金(61402150);中国博士后科学基金资助项目(2016M592286);河南省科技研发专项资助项目(182102410063);三峡大学水电工程智能视觉监测湖北省重点实验室开放基金(2016KLA04).

作者简介(通信作者):李征(1984-),女,河南驻马店人,河南大学副教授,博士,CCF 会员(E200027660M),研究方向为 Web 服务发现与推荐,软件工程,E-mail:lizheng@henu.edu.cn.

进行语义丰富;文献[12]通过词汇间相关度与语义相似度计算获取概念图,然后对图中每个点的度数进行排序以得到概念间的层次关系,但是概念图的构建依赖维基百科与字典;文献[13]根据贝叶斯分类原理利用多个分类器对概念集进行分类,然后通过概念关联分析和概念自学习算法构建概念间的关系,生成本体原型,但是方法中初始分类的个数和术语依靠人工选取.本文利用 K-means 算法自顶向下构建概念间的层次关系,相关参数由相应实验结果确定.

1.基于关联规则和改进 K-means 的领域本体构建方法

初始领域本体构建过程如图 1 所示,首先将 Pweb 网站发布的 Web 服务信息利用爬虫技术保存到本地,然后进行如下操作.

(1)文档预处理:利用自然语言处理工具(NLP)与 Wordnet 进行预处理,得到词频统计文档集;

(2)领域概念获取:针对预处理后的文档集,使用 SVM 对服务进行面向领域的迭代式分类,得到分类后的领域服务集和相应的领域词汇排序表(具体可参考文献[14]),根据服务分类得到的领域词汇排序表,从中选取前 h 个词汇作为初始领域概念;

(3)领域概念间关系挖掘:从初始领域概念中选择种子概念,计算剩余概念与种子概念的置信度,进而构建概念向量空间,使用改进的 K-means 算法进行概念聚类,然后识别每组概念的种子概念并将其作为上层种子概念的子结点;对每组概念重复以上步骤,通过迭代得到初始领域本体.

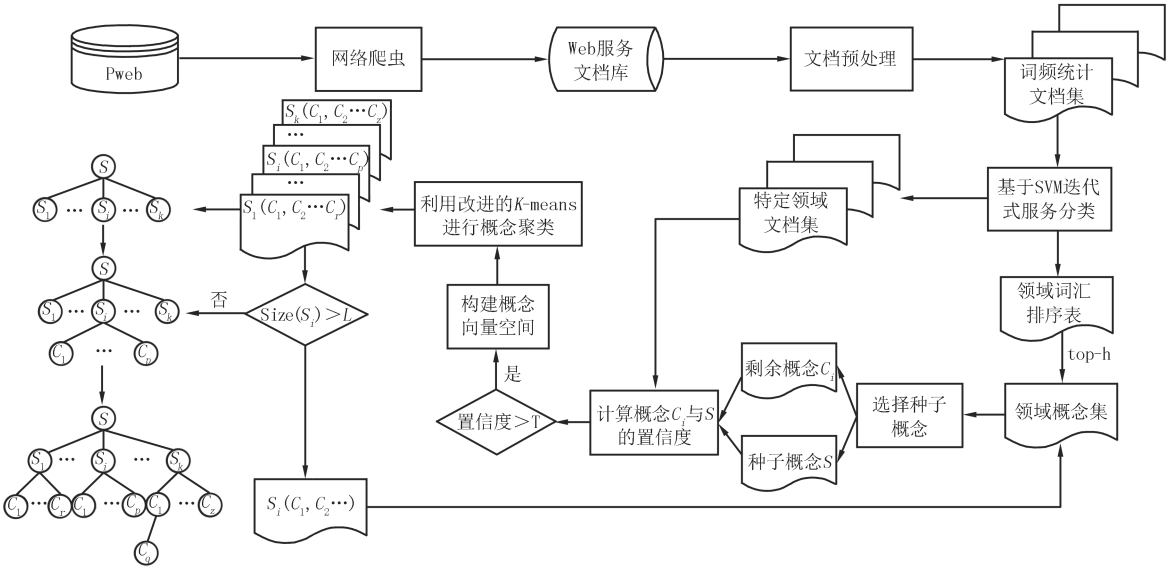

图1　初始领域本体构建过程

Fig.1　Construction process of the initial domain ontology

1.1　文档预处理

针对从 Pweb 上收集的 Web 服务描述文档,首先利用 NLP 工具对其进行如下处理.

(1)分词:利用 NLP 工具解析每个 Web 服务描述文档,获取文档词汇列表.

(2)抽取名词与动词:考虑到名词与动词能够体现 Web 服务功能的核心特征,利用 Wordnet 抽取每个 Web 服务文档词汇列表中的名词与动词,同时,过滤停用词,并利用 Porter Stemming[15] 算法对单词进行词干还原.

(3)统计词频:经过上述步骤后,统计每个词汇在文档中出现的频数,得到相应的 Web 服务词频统计文档集.

1.2　领域概念识别

领域本体提供的服务质量很大程度上取决于领域概念的准确性.在构建领域本体过程中,不仅要获取领域的概念集,而且还要保证概念的准确性.为此,对 1.1 节预处理后得到的服务文档集,使用改进的 $tf\text{-}idf$ 即

$kf\text{-}idf\text{-}df$((1)式),构造向量空间;然后采用支持向量机进行迭代式面向领域的服务分类;对分类得到的领域词汇进行排序,如果领域相关的服务集中连续 2 次迭代得到的领域词汇排序表的前 50 个词汇保持不变,终止迭代.具体可参考文献[14].

$$kf\text{-}idf\text{-}df_{t,w,d}=\begin{cases}tf\text{-}idf_{t,w}\cdot(1+1-\llcorner\dfrac{R(t,d)}{\sqrt{\Omega}}\lrcorner/\sqrt{\Omega})\cdot\mu), & R(t,d)<\Omega,\\ tf\text{-}idf_{t,w}, & R(t,d)>=\Omega,\end{cases}\tag{1}$$

$$tf\text{-}idf_{t,w}=\frac{N(t,w)}{\sum_{t_i\in w}N(t_i,w)}\cdot\lg\frac{\sum_{d_j\in D}|d_j|}{1+|\{w:t\in w\}|}.\tag{2}$$

(1)式中,$kf\text{-}idf\text{-}df_{t,w,d}$表示领域 d 中,服务文档 w 中的关键词 t 在 w 的向量空间中的权重.$tf\text{-}idf_{t,w}$表示文档 w 中的关键词 t 的词频-逆向文档频率,使用(2)式计算得到;$R(t,d)$表示文档 w 中的关键词 t 在文档 w 所属的领域 d 中的词汇排名,即关键词 t 在分类后得到的领域词汇排序表中的位置,排名越靠前,则关键词 t 对领域的表征能力越强,构建向量空间时该词的权重应该放大;Ω 表示领域词汇排序表的前 Ω 个关键词,μ 表示系数.

(2)式中,$N(t,w)$表示关键词 t 在服务文档 w 中出现的次数,$\sum_{t_i\in w}N(t_i,w)$表示文档 w 中所有关键词出现的总次数,$\sum_{d_j\in D}|d_j|$表示领域 D 中的文档总数,$|\{w:t\in w)|$表示领域 D 中包含关键词 t 的文档数.

对服务分类得到的领域词汇排序表,选取前 50 个词汇作为初始领域概念集,为概念间关系的抽取奠定基础.

1.3 领域概念间关系的抽取

本节根据得到的初始领域概念集,采用关联规则和改进的 K-means 算法挖掘概念间的层次关系(纵向关系)与非层次关系(横向关系),通过迭代从上到下逐步构建初始领域本体,具体描述如下.

1.3.1 种子概念识别

领域本体中的种子(核心)概念往往与其他概念有很强的联系,这种联系通常体现在共现关系上,比如种子概念与其他概念共同出现在同一个 Web 服务文档中.文中通过(3)式选取种子概念,选取的标准是特定领域内与众多概念共现次数之和最大的概念.其中,S 代表种子概念,$F(C_i,C_j)$代表概念 C_i 与 C_j 在所有文档中共同出现的频数之和(由(4)式得到),m 代表初始领域概念数.(4)式中,$f(C_i,n,C_j,n)$表示在第 n 个 Web 服务文档中概念 C_i 与 C_j 共同出现的频数,N 代表特定领域的 Web 服务文档数.

$$S=\arg\max_{1\le i\le m}\{\sum_{j=1且j\ne i}^{m}F(C_i,C_j)\},\tag{3}$$

$$F(C_i,C_j)=\sum_{n=1}^{N}f(C_{i,n},C_{j,n}).\tag{4}$$

下面以表 1 为例阐述种子概念的识别过程.在表 1 中,假设 travel 与 book 在所有 Web 服务文档中共同出现次数为 8(由(4)式得到),与概念 hotel 共同出现的次数为 14,与概念 search 共同出现的次数为 17,即概念 travel 与其他概念共现的次数之和为 39.根据(3)式以此类推,travel 与其他概念共现次数之和最大,因此选择 travel 作为最上层种子概念.

1.3.2 概念间层次(纵向)关系挖掘

一般情况下,具有较高支持度与置信度的概念间有很强的关联关系.基于此,利用关联规则中的支持度与置信度来挖掘概念间的层次关系.(5)式与(6)式分别表示支持度与置信度的计算方法.

$$S(C_i)=F(C_i,C_j)/n,\tag{5}$$

$$C_f(C_i)=F(C_i,C_j)/F(C_i).\tag{6}$$

(5)、(6)式中,$F(C_i,C_j)$表示概念 C_i 与 C_j 在特定服务文档中的共现频数;$F(C_i)$表示概念 C_i 在领域中出现的频数;n 表示特定领域内所有概念出现的频数之和.支持度 $S(C_i)$与置信度 $C_f(C_i)$越大,则关联关系越

强.因此,本文中将初始种子概念作为初始领域本体的根结点,将与初始种子概念置信度大于阈值 T 的概念作为该概念的子结点.

以图 2 为例,假设概念 C_1,C_4 与初始种子概念 S 的置信度大于阈值 T,表明它们之间有很强的关联关系,即概念 C_1,C_4 总是伴随概念 S 的出现,那么概念 C_1,C_4 作为概念 S 的子结点.同时,概念 C_2 的出现总是在种子概念与 C_1 出现的前提下,那么概念 C_2 作为概念 C_1 的子结点.以此类推,来构建概念间的层次关系.

表 1　概念共现次数

Tab.1　Number of concept co-occurrences

	travel	book	hotel	search	sum
travel	/	8	14	17	39
book	8	/	10	8	26
hotel	14	10	/	9	33
search	17	8	9	/	34

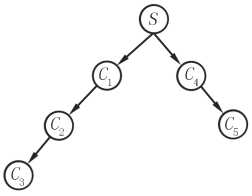

图2　概念间层次关系构建

Fig.2　Hierarchical relationship building among concepts

1.3.3　概念间非层次(横向)关系挖掘

在初始领域本体构建过程中,本文采用改进的 K-means 算法(算法 1)对概念进行聚类,以确定种子概念的子结点数.首先利用(3)式从初始领域概念集中选取初始种子概念,由剩余概念集中每个概念与初始种子概念的置信度和 $kf\text{-}idf\text{-}df$ 构建概念向量集,然后利用改进 K-means 进行聚类,将聚类得到的概念集继续进行上述操作,以此类推迭代式挖掘领域本体中概念间层次关系,其过程如图 3 所示.

图 3 中,(a)表示初始种子概念 S 与普通概念集$\{C_1,\cdots,C_8\}$,S 的获取如 1.3.1 节所述;(b)表示基于 S 的概念聚类,首先根据普通概念与 S 的置信度以及它们各自的 $kf\text{-}idf\text{-}df$ 值构建向量空间,然后利用算法 1 进行聚类,得到 K 组概念集;(c)表示概念分类后根据(3)、(4)式得到每组的种子概念;(d)表示将初始种子概念 S 与每组的种子概念相连接;(e)表示对聚类后的每组概念进一步进行同样处理,迭代后形成的概念间层次关系.

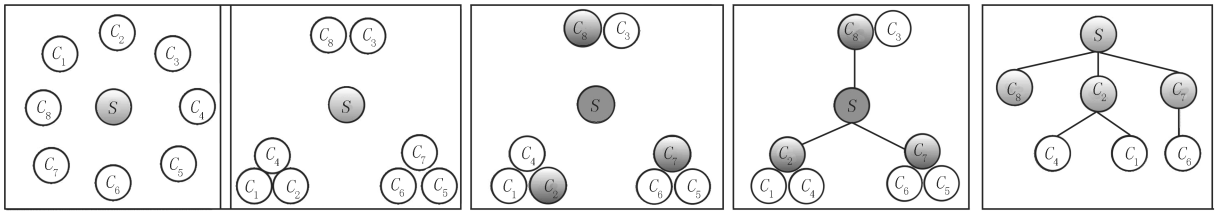

图3　概念间关系挖掘过程

Fig.3　Mining process of relationships among concepts

本文中,利用改进的 K-means 算法进行聚类.首先需要定义一个向量 $\vec{m}=(C_f(S_j,C_i),W(C_i))$.其中,$C_f(S_j,C_i)$ 表示概念 C_i 与种子概念 S_j 间的置信度,$W(C_i)$ 表示概念 C_i 在特定领域内 Web 服务描述文档中的权重(由(1) 式得到);然后随机选取某个概念点作为基准点$(C_f(S_j,C_j),W(C_j))$,遍历其余各概念点$(C_f(S_j,C_i),W(C_i))$ 与基准点之间的欧式距离 d_{ij},如果 $d_{ij}<L_e$,将该点与基准点合并在同一组,反之将该点放入新的一组集合中,得到 K 组概念,进而找到各组概念集的聚类中心;进一步将 K 与各聚类中心作为初始条件,对构建的其他概念与种子概念的向量集进行 K-means 聚类.其改进之处在于可以自动确定聚类的个数 K 与初始点的选择,具体如算法 1 所示.

在算法 1 中,步骤 1)至步骤 8)通过随机选择一个点 $V_i(C_f(S_j,C_i),W(C_i))$作为一个簇,然后计算其他概念 C_j 与 C_i 的距离 d_{ij},如果 d_{ij} 大于阈值 L_e,将 C_j 作为一个新的簇,否则,将 C_i 与 C_j 合并,以获得聚类个数 K 与聚类中心;步骤 9)至 10)利用聚类个数 K 与聚类中心作为初始条件对概念进行 K-means 聚类.

算法 1:改进的 K-means 算法

输入:概念间距离阈值 L_e,概念数据点集合 $V=\{V_i|V_i=(C_f(S_j,C_i),W(C_i)),i=1,\cdots,N\}$;

输出:聚类后 K 个概念簇 $C=\{S_1,S_2,\cdots,S_k\}$.

1)概念数据点集合 $P=V$;//P 用来聚类,V 用来寻找初始聚类中心;

2)当 V 不空;

3)任意选择某个点 V_j 作为初始簇 $U_1=\{V_j|V_j\in V\}$,同时从 V 中移除 V_j;

4)根据 $V_i(C_f(S_j,C_i),W(C_i))$,计算 $V_i\in V$ 与 $V_j(C_f(S_j,C_j),W(C_j))$之间的欧式距离 d_{ij};

5)如果 $d_{ij}<L_e$,将 V_i 添加到 U_1 中,同时从 V 中移除 V_i;

6)否则,选取下一个点,重复步骤 2)至步骤 5);

7)得到初始簇集合 $U=\{U_1,U_2,\cdots,U_k\}$;

8)选择 U 的 k 个簇的质心 $U_{c1},U_{c2},\cdots,U_{ck}$作为初始点;

9)将 P 中每个点指派到最近的质心 $U_{ci}(i=1,2,\cdots,k)$,形成 K 个簇;

10)重新计算每个簇的质心,直到簇的质心不发生变化,迭代终止.

初始领域本体构建如算法 2 所示.其中,步骤 1)至步骤 6)根据(3)、(4)式选择种子概念,步骤 7)至步骤 8)根据置信度与 $W(C_i)$构建概念向量集,步骤 9)利用算法 1 进行概念聚类,步骤 10)至步骤 12)根据聚类后的结果进行迭代处理,步骤 13)输出初始领域本体.

算法 2:初始领域本体构建算法

输入:置信度阈值 T,每组概念数阈值 L,领域概念集 $set=\{C_1,C_2,\cdots,C_n\}$;

输出:初始领域本体.

1)while set.size()$>L$;

2)foreach C_i in set;

3)foreach C_j in set;

4)计算 C_i 与 C_j 的共现频率 $f(C_i,C_j)$;

5)计算 Max{sum($f(C_i,C_j)$)($i,j=1,2,\cdots,n$)};

6)将概念 C_i 作为种子概念,将 C_i,set 放入 Map 中;

7)计算概念 C_i 与 C_j 的置信度 $C_f(C_i,C_j)$;

8)当 $C_f(C_i,C_j)>T$ 构建 $\vec{m}=(C_f(C_i,C_j),W(C_j))$;

9)利用算法 1 对进行聚类得到

$$\text{Clusters}=\{S_1\{C_1,C_2,\cdots C_r\},S_i\{C_1,C_2,\cdots C_p\},\cdots S_k\{C_1,C_2,\cdots C_z\}\};$$

10)foreach $S_i\{C_1,C_2,\cdots C_p\}$ in Clusters

11) set=$S_i\{C_1,C_2,\cdots C_p\}$;

12) 重复以上步骤;

13)输出 Map.

2 实验分析

2.1 实验准备

本文的实验与算法是在 Eclipse 平台下用 Java 语言实现的.所有实验运行在一台具有 AMD Phenom (tm)II X4 B97 Processor 3.20 GHz,4 GB 内存,操作系统为 Window 7 的 PC 上.

实验数据来源于服务网站 Pweb.图 4 所示为 Pweb 上“ALLmyles”API 的相关信息,包括 API 的 name,tags,profile 等.将每个 API 的这些信息以文本文档的形式保存在本地,一共收集了 14 751 个 Web 服务的描述信息.为了验证所提方法的可行性与有效性,本文选取大家熟悉的 Travel 领域的 391 个服务文档进行实验验证,以方便理解文中提出的方法.

2.2　实验评估指标

本文采用准确率(P_r)、召回率(R_e)与 F 值作为评估指标,计算公式如下:

$$P_r=\frac{R_e^c \cap R_i^c}{R_i^c}, \tag{7}$$

$$R_e=\frac{R_e^c \cap R_i^c}{R_e^c}, \tag{8}$$

$$F=2\times\left(\frac{P_r\times R_e}{P_r+R_e}\right), \tag{9}$$

其中,R_e^c 是采用相关方法得到的领域概念集,R_i^c 是人工选取的 Travel 领域标准概念集,具体选取过程如下:首先由 3 名研究生从服务分类得到的 Travel 领域的词汇排序表中选取他们认为能够代表该领域的核心词汇,在此基础上筛选出至少两人都选取的词汇加入概念集,然后由该领域具有丰富经验的领域专家进一步确认决定.

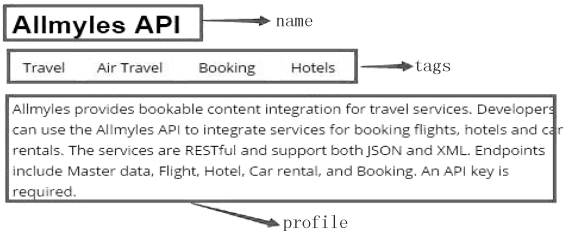

图4　Pweb上“ALLmyles” API的相关信息

Fig.4　Information about the “ALLmyles” API on Pweb

2.3　实验结果分析

本文针对初始领域本体构建的两个重要步骤:即领域概念识别、概念间关系挖掘设计了相关实验,相应实验及结果分析描述如下.

2.3.1　领域概念识别的准确度分析

针对收集的 Travel 领域的服务描述文档,首先根据 1.1 节所述的步骤进行预处理,然后采用支持向量机对预处理后的服务文档进行面向领域的迭代式服务分类,分类后得到 Travel 领域服务文档集和领域词汇排序表,从中选取 top-50 个词汇作为初始领域概念,这里列出 top-10,如表 2 所示($kf\text{-}irf$ 表示该词对 Travel 领域的重要程度,具体参考文献[14]).

表 2　领域概念排序表

Tab.2　Sorting list of domain concepts

概念	travel	booking	hotel	flight	air	airport	reservation	search	rental	airline
$kf\text{-}irf$	0.792 6	0.201 3	0.167 5	0.135 6	0.101 6	0.077 4	0.067 6	0.051 1	0.046 7	0.045 5

为了验证本文提出的概念抽取方法的准确度,采用准确率、召回率、F 值 3 个指标,分别与已有 3 种概念抽取方法进行对比,结果如图 5 所示.

(1)Rank:本文提出的方法.

(2)sum[4]:领域中出现词汇的频数,选取 top-k 个关键词作为领域概念.

(3)textrank[16]:利用 PageRank 进行关键词的抽取.

(4)tfidf:利用传统的 $tf\text{-}idf$ 进行关键词的抽取.

从图 5 可以看出,本文方法从准确率、召回率、F 值 3 个方面都优于其他 3 种方法.进一步分析发现,产生这种结果的原因是:采用(1)式构造向量空间时,不仅考虑了词汇传统的 $tf\text{-}idf$ 值,同时考虑了词汇对领域的表征程度,并且从面向领域的迭代式服务分类产生的领域知识中挖掘领域概念集.

2.3.2　相关参数对构建概念层次关系的影响

本节阐述相关参数对构建领域本体中概念层次关系的影响.本文采用动态值设定置信度的阈值,因为置信度越小获取的概念数越多,但结果往往不令人满意,而置信度越大获取的概念数越少,不利于领域本体的表示,因而通过置信度均值与标准差的差值选取最佳置信度.

图 6 中,M 代表置信度的均值,D 代表置信度的标准差.M-D、M-2D、M-3D 分别表示置信度的均值 M 与

1 倍、2 倍、3 倍标准差的差值,并将其分别设置为置信度阈值 T,相应准确率、召回率与 F 值的结果如图 6 所示.

从图 6 可以看出,置信度的取值不同,准确率、召回率与 F 值的结果也不同,3 个值随着置信度的减小而增加,当置信度取值为 M-2D 时,3 个值均达到最大,然后随着置信度的减小而减小.因此,实验中取置信度为 M-2D 进行概念间关系构建.

类似的,算法 2 中每组概念数阈值 L 分别取值为 1,2,…,20 进行实验.结果表明 L 取值为 5 时,实验效果较好.因此,实验中设置 L 值为 5.

在算法 1 中,聚类时概念间距离阈值 L_e 分别取值为所有向量距离均值的 0.1,0.2,…,1.0 倍进行实验.结果表明将 L_e 设置为所有向量距离均值的 0.5 倍时效果最好.因此,实验中 L_e 取值为所有向量平均距离的 0.5 倍.

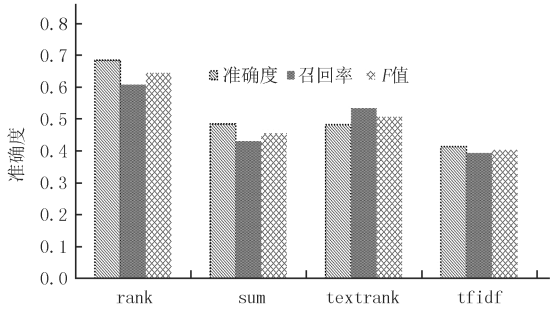

图5 概念识别准确度对比

Fig.5 Comparison of concepts recognition accuracy

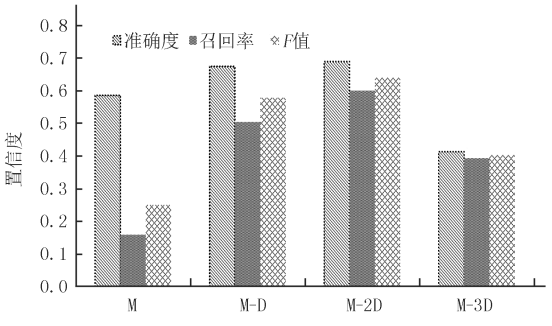

图6 置信度取值对概念层次关系的影响

Fig.6 The Influence of confidence value on conceptual hierarchy

2.3.3 概念间关系挖掘分析

根据 1.3.1 节得到的领域概念集,找到初始种子概念 travel,然后构建基于 travel 的向量集,将基于 travel 的向量集利用算法 1 进行聚类,利用算法 2 对聚类后的各组概念重复上述操作,得到初始领域本体.利用 Protégé 对初始本体结果进行可视化展示,如图 7 所示.

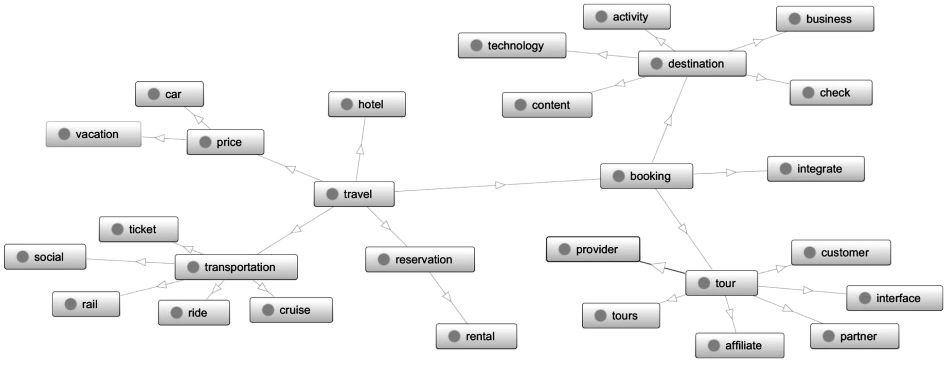

图7 初始领域本体展示

Fig.7 Display of the initial domain ontology

从图 7 可以看出,概念间具有明显的层次关系,可以看到对词汇关联关系较大的概念进行聚类,聚类中心能够对本组概念进行“代表”,不同的中心概念形成不同的“代表”,进而构建出概念间的关系,从而验证了方法的可行性.

2.3.4 利用 Wordnet 对初始领域本体进行扩充

领域本体中的概念关系一般包括同义关系、上下位关系等,对初始领域本体中的每个概念,根据 Wordnet 找到其相关关系的词汇进行补充.例如,通过 Wordnet 得到的 hotel 的下位词有 fleabag,hostel,hostelry

等,car 的同义词有 auto,automobile 等.类似地,利用 Wordnet 得到的词对初始领域本体中的相应概念进一步丰富,结果如图 8 所示.从图 8 可以看出:通过 Wordnet 可以很好地丰富初始领域本体的语义,进而为语义 Web 服务查询提供相应支持.

3　结　论

本文提出一种基于关联规则和改进 K-means 的领域本体构建方法,包括数据收集、预处理、迭代式的服务分类、初始领域概念识别及概念间关系构建等.最后,以 Pweb 上真实的服务集进行实验,验证了所提方法的可行性和有效性.但是,论文仍存在以下不足:(1)领域中的概念多是单个词汇,缺少以多个词汇出现的概念;(2)概念间的语义关系不够丰富,构建出的概念间层次关系需要进一步评估.下一步工作将重点围绕上述两点展开,具体包括增加多词汇概念,利用 Wordnet 与 Word2vec 进一步丰富概念间的上下位关系、整体与部分关系等,同时对概念间关系进行评估.

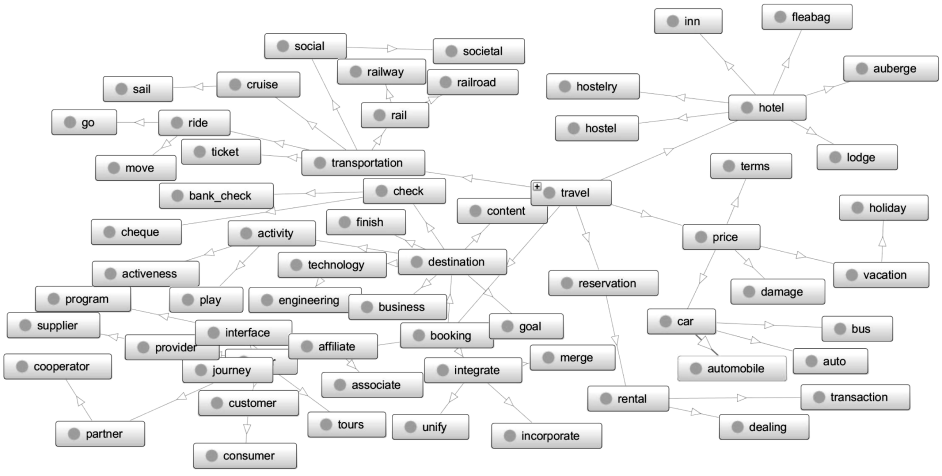

图8　初始领域本体概念的语义丰富

Fig.8　Semantic enrichment of the initial domain ontology concept

参　考　文　献

[1]　GRUBER T R.A translation approach to portable ontology specifications[J].Knowledge acquisition,1993,5(2):199-220.

[2]　DENG Z,TANG S,ZHANG M.Overview of ontology[J].ACTA Scientiarum Naturalium-Universitatis Pekinensis,2002,38(5):730-738.

[3]　周子力.基于 Wordnet 的本体构建及其在安全领域应用关键技术研究[D].上海:华东师范大学,2009.
ZHOU Z L.Research on key technologies of ontology construction based on WordNet and its application in security domain[D].Shanghai:East China Normal University,2009.

[4]　Küçük D, ARSLAN Y. Semi—automatic construction of a domain ontology for wind energy using Wikipedia articles[J].Renewable Energy,2014,62:484-489.

[5]　LIU B S,GAO J.General ontology learning framework[J].Journal of Southeast University(English Edition),2006,22(3):381-384.

[6]　陈刚,陆汝钤,金芝.基于领域知识重用的虚拟领域本体构造[J].软件学报,2003,14(3):350-355.
CHEN G,LU R Q,JIN Z.Constructing virtual domain ontologies based on domain knowledge reuse[J].Journal of Software,2003,14(3):350-355.

[7]　LEE C S,KAO Y F,KUO Y H.Automated ontology construction for unstructured text documents[J].Data & Knowledge Engineering,2007,60(3):547-566.

[8]　DONG X,HALEVY A,MADHAVAN J,et al.Similarity search for web services[C]//Proceedings of the Thirtieth international conference on Very Large Data Bases-Volume 30.[s.l.:s.n.],2004:372-383.

[9]　韩道军,甘甜,叶曼曼,等.基于形式概念分析的本体构建方法研究[J].计算机工程,2016,42(2):300-306.
HAN D J,GAN T,YE M M,et al.Research on ontology construction method based on formal concept analysis[J].Computer Engineering,2016,42(2):300-306.

[10]　TA C D,THI T P.Improving the formal concept analysis algorithm to construct domain ontology[C]//Proceedings of 2012 Fourth inter-

national conference on Knowledge and Systems Engineering.[s.l.:s.n.],2012:74-78.

[11] 田刚,何克清,孙承爱,等.Web 服务描述的本体学习方法[J].计算机科学与探索,2015(5):575-585.
TIAN G,HE K Q,SUN C A.Ontology learning from Web service descriptions[J].Journal of Frontiers of Computer Science and Technology,2015(5):575-585.

[12] YU H,LYU X Q,XU L P.Use Web resources to construct ontology concept hierarchy[C]//Proceedings of the international conference on Applied Science and Engineering Innovation.Paris:Atlantis Press,2015:1006-1011.

[13] 金鑫.面向 Web 信息资源的领域本体模型自动构建机制的研究[J].计算机科学,2012(6):213-216.
JIN X.Research on mechanism of automatic construction of ontologies for Web information resources[J].Computer Science,2012(6):213-216.

[14] WANG J,ZHANG J,Hung P C K,et al.Leveraging fragmental semantic data to enhance services discovery[C]//Proceedings of 2011 IEEE international conference on High Performance Computing and Communications.[s.l.:s.n.],2011:687-694.

[15] PORTER M.An algorithm for suffix stripping[J].Program:electronic library and information systems,1980,14(3):130-137.

[16] 杨洁,季铎,蔡东风,等.基于 TextRank 的多文档关键词抽取技术[C]//第四届全国信息检索与内容安全学术会议论文集(上).北京:[出版者不详],2008:397-404.
YANG J,JI D,CAI D F,et al.Keyword extraction in multi-document based on TextRank technology[C]//Proceedings of the 4th national conference on Information Retrieval and Content Security.Beijing:[s.n.],2008:404-411.

An approach for domain ontology construction based on association rules and *K*-means

Li Zheng[1,2],Li Bin[1]

(1.School of Computer and Information Engineering,Henan University,Kaifeng 475004,China;2.Key Laboratory of Intelligent Vision Monitoring for Hydropower Project of Hubei Province,China Three Gorges University,Yichang 443002,China)

Abstract:With the scale growth of service resources on the network,how to help users find their required services is a key issue.Studies found that the construction of domain ontology can help users effectively solve the problem,but some of existing methods are built manually,which is time-consuming and laborious.In order to solve this problem,this paper proposed an ontology construction method based on association rules and *K*-means.Firstly,we use support vector machine to conduct domain-oriented services classification,and select initial domain concepts from the domain knowledge obtained by classification.Then according to association rules and *K*-means,we mine the relationships among concepts to obtain the initial domain ontology,and the obtained ontology is further enriched by Wordnet.Finally,the real services from ProgrammableWeb are used to conduct experiments.The experimental results show that the proposed domain ontology construction approach can identify the relationships among concepts,and then provide the corresponding support for Web services semantic query.

Keywords:service classification;association rule;*K*-means;domain ontology

[责任编校 陈留院 赵晓华]