第 40 卷 第 12 期 计 算 机 工 程 2014 年 12 月
Vol. 40 No. 12 Computer Engineering December 2014

·先进计算与数据处理· 文章编号:1000-3428(2014)12-0057-06 文献标识码:A 中图分类号:TP391

一种混合的领域概念分类体系自动构建算法

罗年洁,吕 钊

(华东师范大学计算机科学技术系,上海 200241)

摘 要: 领域概念分类体系自动构建在人工智能、自然语言处理和信息检索等领域具有重要作用,但现有研究较多关注通用知识,面向特定领域的研究较少,且存在领域概念间关系抽取准确率以及自动构建算法效率较低等问题。为此,提出一种混合的领域概念分类体系自动构建算法,该算法主要包括领域概念间关系抽取模块和分类体系构建模块。领域概念间关系抽取模块设计考虑中文自身的特点,采取句法树和基于规则相结合的方法,以提高抽取领域概念间关系的查准率和查全率;分类体系构建模块设计采取改进的 BRT 算法,从而在降低算法复杂度的同时,提高领域分类体系构建的查准率。在通信、金融和计算机领域的实验结果均表明,与 BRT 算法相比,该算法的构建效果较好,查准率最高可达到 89.3% 。

关键词: 领域概念分类体系;贝叶斯玫瑰树;句法树

中文引用格式: 罗年洁,吕 钊. 一种混合的领域概念分类体系自动构建算法[J]. 计算机工程,2014,40(12):57-62,67.

英文引用格式: Luo Nianjie,Lü Zhao. A Hybrid Algorithm of Automatic Domain Concept Taxonomy Construction[J]. Computer Engineering,2014,40(12):57-62,67.

A Hybrid Algorithm of Automatic Domain Concept Taxonomy Construction

LUO Nianjie,LÜ Zhao

(Department of Computer Science and Technology,East China Normal University,Shanghai 200241,China)

【Abstract】 Domain concept taxonomy automatic construction plays an important role in artificial intelligence,natural language processing and information retrieval. Existing approaches pay more attention on common knowledge,while there are fewer reports about domain concepts. Two main challenges of domain concept taxonomy automatic construction are identifying relationships between concepts and less efficiency of current algorithms. In this paper,a Hybrid algorithm of Automatic Domain concept Taxonomy construction(HADT) is proposed,which has two main modules:extracting relationships between domain concepts and automatic taxonomy construction. Considering Chinese characteristics,the first module uses syntax tree method and rule-based method together,to get the aim of higher precision and higher recall. The second module uses an improved BRT algorithm to reduce time complexity and to improve taxonomy construction precision. The experiments conducted on three datasets of mobile,financial and computer show the HADT algorithm is effectiveness compared with the BRT algorithm,and the highest precision rate is 89.3% .

【Key words】 domain concept taxonomy;Bayesian Rose Tree (BRT);syntax tree

DOI: 10.3969/j.issn.1000-3428.2014.12.010

1 概述

领域概念层次结构是对特定领域的概念特征化描述,可反映该领域内的知识和概念关系。它不仅有助于知识推理,促进信息分类、搜索与导航,而且有助于人或机器理解一个高度集中或快速变化的领域[1-2]。一般地,领域概念层次构建主要有 2 个部分:领域概念间关系抽取与层次构建。现有面向领域的概念自动构建方法主要有 2 类:基于知识库的方法和基于原始数据的方法。

基于知识库的方法是通过已有的结构化或半结构化的知识网获取领域词对关系,然后构建概念层

基金项目: 国家科技支撑计划基金资助项目(2012BAH74F02);上海市科委科研基金资助项目(12dz1500205)。

作者简介: 罗年洁(1989-),女,硕士研究生,主研方向:大数据分析,知识处理;吕 钊(通讯作者),副教授。

收稿日期: 2013-12-24 **修回日期:** 2014-01-15 **E-mail:** zlu@cs.ecnu.edu.cn

次。例如,文献[3]使用 WordNet 构建餐饮系统;文献[4]首先从维基百科抽取领域词对关系,然后采用有向无环图(Directed Acyclic Graph,DAG)算法构建领域概念图,最后通过深度遍历来建立层次结构;文献[5]使用其构建的知识库 Probase 获取领域概念关系,再对给定的关键词集使用贝叶斯玫瑰树(Bayesian Rose Tree,BRT)[6]来构建汽车保险领域的层次结构。

基于原始数据的方法主要依赖于纯文本文档,如文献[7]基于频率来统计词对的共现概率,大于一定阈值则具有上下位关系,然后根据其算法(Fuzzy OntoExt)构建概念图;文献[8]基于形式概念分析的研究,首先采用 FCA 算法构建概念格,然后使用 K-Means 算法对概念进行聚类。

基于知识库的方法获取领域概念关系比较方便且准确率高,然而随着领域发展,会更新很多词意,并产生新词,知识库拓展性弱,并未能实时反馈这些改变,这样会导致领域概念间关系的查全率降低[9]。基于原始数据的方法会忽略了低频领域词。FCA 算法比较适用于对象-属性类型的领域,最后得到的是领域的概念格,而不是一个直接的领域层次结构。

国内开展了中文概念层次结构构建研究,如文献[10]构建了中文词典的层次结构,其主要是先定义词典中词的语义框架,取得了较好的效果,但该方法扩展性弱,不易移植到其他领域。文献[11]在获取领域术语后,采用一种自顶向下的聚类算法获取领域概念间的层间关系,这种方法的聚类层数需要人工确定,无法自动获取完整的层次关系。

现有的领域概念层次自动构建方法主要存在以下两方面的问题:(1)领域概念间的关系查全率低;(2)构建算法复杂度高。为此,本文提出一种混合的领域概念层次结构自动构建算法(DCTA)

2 混合的领域概念层次结构自动构建算法

DCTA 算法的主要步骤如图 1 所示。

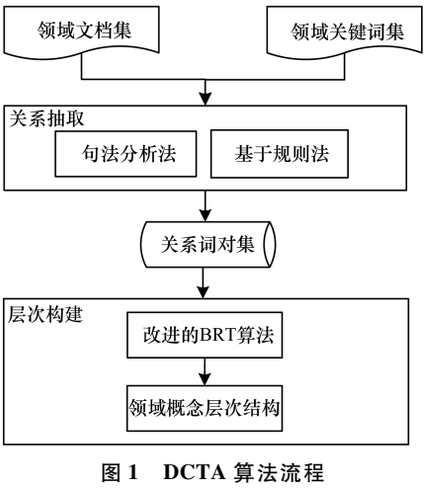

图 1 DCTA 算法流程

本文主要采用全自动方法构建中文领域的分类体系,主要包括关系抽取和层次构建 2 个部分。

领域词对关系的抽取过程如下:(1)输入领域文档集与领域关键词词集,使用“领域词与上下文”的模式,获取领域词对共现的句子;(2)使用句法树和基于规则的方法,获取满足要求的词对集;(3)对以上 2 个方法获取的词对集合进行合并。

层次结构自动构建是对存在关系的领域词构建层次结构,其步骤如下:(1)输入关系词对集,通过改进的 BRT 算法对节点进行合并、吸收和连接操作;(2)将所有的子节点归并到一个父节点下,构建领域层次结构。

2.1 领域概念关系抽取

领域概念是在特定领域文档频繁出现、反映该领域共性的特征词语,具有专指性强、领域区分度高、领域化代表性强的特点。随着信息的发展,领域不断更新,出现了很多新词,其中包括大量的复合词,即多个词组成的多字概念,对于这些新词,已有的知识库未能实时更新,而且复合词根据构成词存在一定领域层次关系。根据领域词对关系的特点,本文使用分治策略对句子中的一些特定语法结构进行预处理,选择使用句法树和基于规则的方法。

由于句法分析是对语言进行深入理解的基础,它从句子结构上分析领域词对关系,具有语料处理快、标注方法和算法先进、标注标准和其他语料库的兼容性较好等优点。采用句法树可以很好地提取领域中词对间关系。

基于规则的方法被广泛应用于关系识别和人名识别[12]等领域,可以最大限度地接近自然语言的句法习惯,从而被快速掌握;其表达方式灵活多样,能最大限度地表达研究人员的思想;同时也能很好地解决复合词包含的领域关系。

2.1.1 句法树分析法

本文采用的句法树是一个词汇化的概率上下文无关文法(Probalilistic Context Free Grammar,PCFG)[13]语法分析器,句法分析模型句法树分析的结果一般表示为树结构,树的节点表示句子的语法单元的名称,而树的分叉表示 2 个或者多个语法单元组成一个新的、跨度更大的语法单元。

例如对“神州行幸福卡是一款专为老年客户设计的具有月费低,亲情号码通话优惠的资费套餐。”进行句法分析,其中,“神州行幸福卡”、“亲情号码”、“资费套餐”是 3 个领域词,从句法树图可以得到领域概念关系“神州行幸福卡”是“资费套餐”,其句法分析树结构如图 2 所示。可以看出,句法树对句子关系能取得很好的结果,但对于名词复合短语

的关系抽取却不理想,如“神州行”与“幸福卡”之间存在着整体与部分关系。名词复合短语是各种语言中普遍存在的一种语法结构,对信息抽取、机器翻译等应用有很大的影响,由于句法分析对此类结构的处理不够理想,本文对名词复合短语进行专门处理,以降低句法分析的难度。针对汉语名词复合短语的特点,提出一种基于规则的名词复合短语分析方法,以减小此类短语对句法分析的影响。

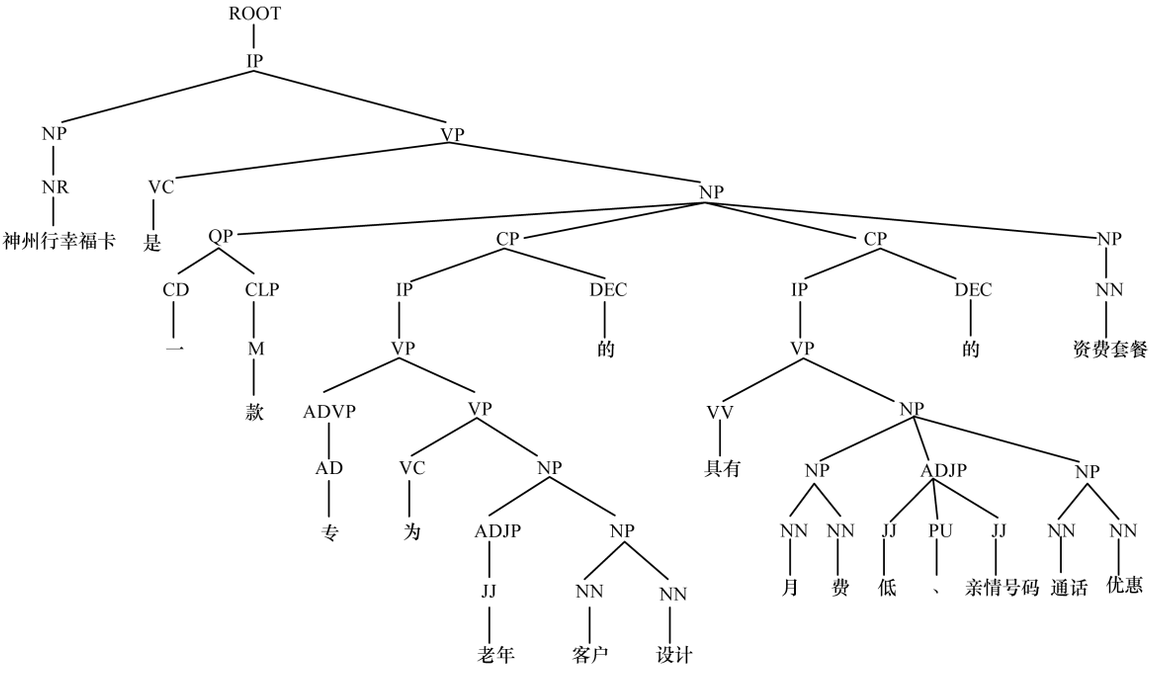

图 2　句法分析实例

2.1.2　基于规则的分析法

为了实现对复合短语与其他复杂句式中领域概念关系的识别,本文引入基于规则的分析法,规则匹配用于发现一些文档中没有明确说明的关系。随着领域的不断拓展,出现了很多复合词,而且这些复合词往往单独在一些语句出现,并未能表示它与领域词的关系,它往往是多个名词的组合形成。通过统计一些名词出现在领域概念的开始或结尾的频率,高频词成为领域词特定的前缀或后缀。根据语义,这些前缀或后缀词与复合词具有上下位关系。比如在手机电信领域中,短信业务、彩铃业务等都含有相同的后缀词业务,这些后缀词业务表达了领域概念“短信业务”、“彩铃业务”的特性与所属类别。本文根据领域词的关系特点制定规则,如表 1 所示。

表 1　规则示例

序号	规则描述	规则例句
1	词后缀模式,例如:＊＊＊业务	主叫彩铃业务可以决定自己拨打别人电话时听什么声音
2	词前缀模式,例如:动感地带＊＊＊	向您推荐动感地带新网聊系列套餐,该系列套餐最高含 600 条国内短信、200 MB 国内手机上网流量,超过后国内上网流量 0.01 元/10 KB,网内短信 0.06 元/条、网外短信 0.1 元/条
3	包括	手机银行的 e 贷通包括产品介绍、办理指南、贷款查询、授信查询及放款、贷款还款。登录使用此菜单需要正式注册用户权限
4	包含	外汇宝-账户管理包含活期查询、合并转期、定期查询、活期转定期、定期转活期几项功能
5	由…组成	积分/M 值主要由消费积分、品牌积分、网龄积分和奖励积分四部分组成

领域概念包含 4 种关系:ISA,Part-Of,Attribute-Of 和 Instance-Of。ISA 关系表示类别之间有共同的属性,用来表示概念的逐步细化,类似于面向对象中的继承概念;Part-Of 关系表示类别之间是整体和部分的关系;Attribute-Of 表示关系表示某对象是一概念的属性;Instance-Of 关系表示某对象是一概念的实例。表 2 为上述 4 种关系类型的关系举例。

表 2　领域词 4 种关系类型举例

序号	领域词 A	领域词 B	领域词对关系类型
1	神州行幸福卡	资费套餐	ISA
2	品牌积分	积分/M 值	Part-Of
3	改号语音通知	语音通知	Attribute-Of
4	动感地带	套餐	Instance-Of

2.2 基于改进的 BRT 领域概念层次构建

本文基于关系抽取获取的领域词对集-构建领域层次。在初始化时每个数据点都是一棵树，如 $T_i=\{x_i\}$，x_i 表示第 i 个数据节点的特征向量，每一步选取 2 个层次结构 T_i，T_j 融合成一个新分类结构 T_m，如图 3 所示。

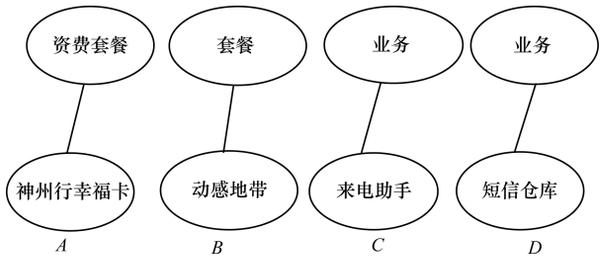

图 3 节点初始化分类体系

本文通过算法比较，确定 2 个层次结构的融合方式，有 3 种融合方式：连接，吸收，归并。

（1）连接（Join）

$T_m=\{T_i,T_j\}$

$leaves(T_m)=leaves(T_i)\cup leaves(T_j)$

其中，*leaves* 表示每个层次结构的所有的叶节点。这个方法是从根节点合并 2 个层次结构，T_m 有 2 个子节点，如图 4 所示。

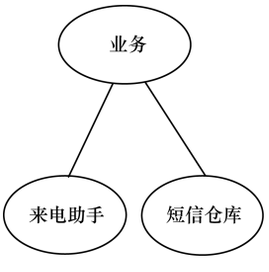

图 4 *C* 节点和 *D* 节点的连接操作

（2）吸收（Absorb）

$T_m=\{children(T_i),T_j\}$

如图 5 所示，此方法可以理解 T_j 变为 T_i 的子节点，对于 2 棵层次结构，也有可能 T_i 变为 T_j 的子节点，此方法逆向的表示为：

$T_m=\{T_i,children(T_j)\}$

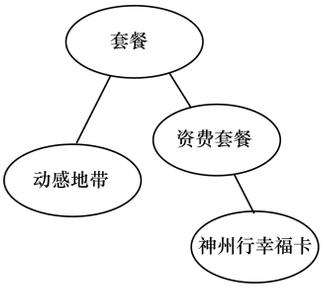

图 5 *A* 节点与 *B* 节点的吸收操作

（3）归并（Collapse）

$T_m=\{children(T_i),children(T_j)\}$

2 棵层次结构的子节点归并成一个层次结构下，如图 6 所示。

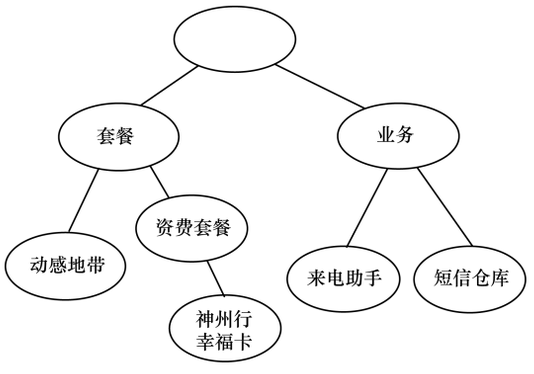

图 6 归并操作

算法每一步挑选 2 个层次结构进行 3 种可能性合并操作，得到每种合并方法的概率，其计算公式如下：

$$p(D\mid T)=\pi_T f(D)+(1-\pi_T)\times\prod_{T_i\in ch(T)}p(leaves(T_i)\mid T_i)$$

其中，D 表示层次结构 T 的所有数据节点；$f(D)$ 代表边缘概率；π_T 是表示所有在 T 的数据不被分成子层次结构的先验概率，其定义如下：

$$\pi_T=1-(1-\gamma)^{n-1}$$

其中，γ 是介于 0 和 1 之间的控制模型超级参数；n 表示 T 中子节点的数目。不同的 γ 对实验结果有很大影响。

对于边缘概率 $f(D)$ 的表示采用了基于多项式的 Dirichlet 共轭分布（DCM）[14]，因此，它更能代表一个或多个主题，在分层聚类中，逐步合并簇。

$$f_{\mathrm{DCM}}(D)=\prod_i^n\frac{m!}{\prod_j^V x_i^{(j)}!}R(x_i,x_j)$$

其中，V 表示词量；$x^{(j)}$ 是相对于 $V_i^{(j)}$ 的频率；$m=\sum_j^V x^{(j)}$；$R(x_i,x_j)=\frac{1}{\sqrt{2\pi\sigma}}\mathrm{e}^{|x_i-x_j|^2}$ 表示第 i 个节点和 j 个节点的关系概率。

对 3 种融合方式，采用比率评分 $Score(D\mid T)$ 形式决定选择哪种融合方式，其公式如下：

$$Score(D\mid T)=\frac{p(D_m\mid T_m)}{p(D_i\mid T_i)p(D_j\mid T_j)}$$

领域层次结构自动构建算法如下：

输入 领域词集

输出 领域概念层次结构

初始化：T{T_i（i = 1，2，…，n）；} count = n

1. For each T_i in T

2. Set value initial（T_i）= f_{DCM}（D）；//赋值

3. End for
4. For each (T_i,T_j),has 3 methods to merge
   D_0 = Absorb(T_i,T_j),D_1 = Join(T_i,T_j)
   D_2 = Collapse(T_i,T_j)
5. Select max(Score(D|T))
6. End for
7. If T_i,T_j has been merged
8. Count--;
9. End if

在原来的 BRT 算法复杂度为 $O(n^2C_v + n^2\log n)$,空间复杂度为 $O(n^2)$,其中,C_v 为所有初始化向量 $\boldsymbol{x}_i$ 中所有非零元素的最大数,当领域词的个数比较多时 C_v 是一个不容忽视的数。本文将领域词对间关系也考虑在内,计算了 2 个领域词关系的概率,降低了了边缘概率的复杂度,此算法将复杂度降低到 $O(n^2\log n)$。其中,领域词对关系为:

$$R(x_i,x_j) = \frac{1}{\sqrt{2\pi\sigma}}e^{|x_i-x_j|^2}$$

如果 2 个节点有关系,更能准确计算 2 个节点的融合方式,将 $R(x_i,x_j)$ 代入节点融合运算,不仅降低了算法复杂度,而且有利于保证层次结构构建的正确性。

3　实验与结果分析

3.1　测试数据集和实验方案

本文实验分别对通信、金融和计算机 3 个领域实现概念层次自动构建。实验评估标准参考文献[15]提出的本体评价标准,其公式如下:

$$P(Ref,Comp) = \frac{|Ref \cap Comp|}{|Comp|}$$

$$R(Ref,Comp) = \frac{|Ref \cap Comp|}{|Ref|}$$

$$F(Ref,Comp) = \frac{2 \times P(Ref,Comp) \times R(Ref,Comp)}{P(Ref,Comp) + R(Ref,Comp)}$$

其中,*Ref* 表示参考的层次结构;*Comp* 表示要比较层次结构。通过上式可以计算领域概念层次结构的查准率 *P*、查全率 *R* 和综合指标 *F* 值。*Ref* 以手动构建的概念层次结构作为参考标准。

关系抽取部分采用 PCFG 的句法结构训练得到的句法分析器进行句法分析,然后采用规则匹配的方法,并将两者结合,统计 ISA 和 Part-of 两种关系的种类,实验表明该 2 种关系占总数的 80%,表 3 展示了 30 对领域概念关系抽取结果。该表是对 3 个领域的词对关系举例说明,部分领域词对间关系如表 2 所示。实验的机器配置如下:处理器为 Intel ® Pentium ® CPU G630 @ 2.7 GHz,内存 4 GB,操作系统为 Win7 64 位。

表 3　领域词对关系

序号	领域词 A	领域词 B	关系
1	神州行幸福卡	资费套餐	ISA
2	亲情家园 A 计划	融合套餐	ISA
3	139 邮箱	电子邮箱	ISA
4	飞信	聊天工具	Part-Of
5	飞信 QQ	增值功能	ISA
6	短信仓库	业务	Part-Of
7	国际短信	业务	Part-Of
8	来电助手业务	业务	Part-Of
9	音乐语音搜索	业务	Part-Of
10	保密天使	需求	ISA
11	改号语音通知	语音通知	Attribute-Of
12	呼叫等待	提示音	Attribute-Of
13	无线音乐语音搜索服务	服务	ISA
14	彩铃音乐盒	业务	Part-Of
15	无线音乐俱乐部	产品	ISA
16	两城一家	增值套餐	Part-Of
17	动感地带	套餐	Instance-Of
18	互联网数据安全	网上银行	Attribute-Of
19	盈利挂盘	外汇宝服务	Part-Of
20	天添利业务	理财业务	Part-Of
21	定期定额投资	基金买卖	Part-Of
22	网银积分	积分	Instance-Of
23	跨行转账	转账交易	Part-Of
24	用户公钥	数字证书	Attribute-Of
25	储蓄国债	证券期货	Part-Of
26	事物故障	故障	Instance-Of
27	数据库管理	数据库系统	Part-Of
28	数据库操纵语言	结构化查询语言	ISA
29	排它锁	意向锁	Part-Of
30	启发式优化	查询优化	ISA

3.2　结果分析

本文实验主要包括 2 个步骤:领域词关系抽取和自动构建层次结构,关系抽取对第 2 步的层次构建有很大影响,因此,本文对两部分进行了实验分析。

3.2.1　领域词对关系抽取的效果评估

该部分主要是获取领域关系,表 4 显示的是分别采用句法树、基于规则的方法,句法树与基于规则结合的方法和人工 4 种方法获取通信领域概念词对关系的对比结果。

表 4　关系抽取效果比较

方法	查找到的关系个数	查准率/%	查全率/%
句法树	220	90.0	51.2
基于规则	356	98.3	90.4
句法树 + 规则	363	98.6	92.5
人工	387	100.0	100.0

中文表达比较模棱两可,再加上句法结构复杂,其中还有一小部分没有识别出来,对其进行分析,其原因如下:

(1) 若 2 个领域概念多次共现在同一句子,共现频率比较高则可能存在一定的关系,本文缺少对此方面的考虑;

(2) 网络抓取的部分句子结构比较长,句法比较复杂,中间干扰词比较多,而结果错误的判给其他领域词。

3.2.2 层次结构自动构建的效果评估

基于上一步获取领域词以及关系,在使用改进的 BRT 算法时,通过不断实验发现当 $\gamma = 0.3$ 时,效果最优。对通信领域数据进行 5 次实验,结果如表 5 所示,其中,n 表示领域词的个数。

表 5 通信领域实验结果 %

n	查准率	查全率	F 值
20	76.9	83.3	80.0
50	80.0	88.9	84.2
100	89.3	92.5	90.1
200	85.0	89.4	87.3
500	88.0	90.1	89.3
平均	83.8	88.9	86.3

笔者发现领域概念数量越多其查准率越高,因为随着领域词丰富,一方面能更多的发现领域词关系,对关系识别越有利;另一方面在层次自动构建上,随着领域概念数量的增多,能更准确地定位到节点,更有利于提高节点融合的准确率。

采用 BRT 算法和改进的 BRT 算法在通信领域的领域概念数 $n = 500$ 时,做了对比实验,其结果如表 6 所示。可以看出,DCTA 算法构建的层次结构查准率最高达 88% ,比使用 BRT 算法提高了 5.4% ,查全率提高了 5% 。实验表明本文算法的可行性。

表 6 2 种构建算法的结果比较 %

算法	查准率	查全率
BRT 算法	82.6	85.1
DCTA 算法	88.0	90.1

另外,将此算法对金融和计算机领域构建层次结构,其结果如表 7 所示。可以看出,本文方法具有很强的移植性,可以适用金融领域和计算机领域。证明该算法充分考虑了领域概念的特点,选择算法可用性强,适用于构建复杂领域层次结构。

表 7 金融领域和计算机领域实验结果 %

领域	查准率	查全率	F 值
金融	83.3	92.5	87.7
计算机	77.8	94.5	85.4

4 结束语

本文提出基于中文面向领域的概念层次自动构建算法(DCTA),该算法主要包括领域词对关系抽取和自动层次构建,即采用句法树和基于规则的混合方法获取领域词关系,然后使用改进的 BRT 算法构建层次结构。本文研究实现了对 3 个领域的概念层次自动构建,并在通信领域与传统的 BRT 算法进行了对比实验。实验结果表明,本文算法具有较高的查准率,而且可移植性强。下一步将改进算法以提高分类体系的准确度,并针对更多的领域进行实验,推导出适用多个相关领域的分类体系算法。

参考文献

[1] Sadikov E, Madhavan J, Wang Lu, et al. Clustering Query Refinements by User Intent[C]//Proceedings of the 19th International Conference on World Wide Web. New York, USA: ACM Press, 2010: 841-850.

[2] Perry P, Wise W, O'Neill D, et al. Leveraging a Technical Domain Taxonomy to Enhance Collaboration, Knowledge Sharing and Operational Support [C]// Proceedings of SPE Digital Energy Conference and Exhibition. The Woodlands, USA: Society of Petroleum Engineers, 2011: 19-21.

[3] Agirre E, de Lacalle O. Publicly Available Topic Signatures for All WordNet Nominal Senses [C]// Proceedings of LREC'04. Lisbon, Portugal: European Language Resources Association, 2004: 1123-1126.

[4] Deshpande O, Lamba D S, Tourn M, et al. Building, Maintaining, and Using Knowledge Bases: A Report from the Trenches [C]//Proceedings of 2013 International Conference on Management of Data. New York, USA: ACM Press, 2013: 1209-1220.

[5] Liu Xueqing, Song Yangqiu, Liu Shixia, et al. Automatic Taxonomy Construction from Keywords [C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2012: 1433-1441.

[6] Blundell C, Teh Y W, Heller K A. Bayesian Rose Trees[C]//Proceedings of UAI'10. Catalina Island, USA: [s. n.], 2010: 68-76.

[7] Monachesi P, Markus T, Mossel E. Ontology Enrichment with Social Tags for eLearning[C]//Proceedings of the 4th European Conference on Technology Enhanced Learning. Nice, France: [s. n.], 2010: 385-390.

[8] Lupea M, Tatar D, Marian Z. Learning Taxonomy for Text Segmentation by Formal Concept Analysis[C]// Proceedings of CoRR'10. [S. l.]: Springer, 2010: 84-92.

[9] Tsui E, Wang W M, Cheung C F, et al. A Concept-relationship Acquisition and Inference Approach for Hierarchical Taxonomy Construction from Tags [J]. Information Processing & Management, 2010, 46(1): 44-57.

(下转第 67 页)

大小,*rate* 值设置的越大,说明对经停地需要的密度分布越大,那么找到的经停地就越小和越少。在使用该参数时,要考虑到 th_{imco} 的影响,如果 *rate* 选择的太小,导致找到的经停地太小,那么在经停地内就可能不会出现持续时间超过 th_{imco} 的轨迹,导致找不到重要同现;如果 *rate* 取得太大,会使得找到的时空点太多,达不到限制区域和减少冗余计算的目的。

5　结束语

本文通过分析青海湖斑头雁的数据可知,该同现模式算法挖掘出的时空模式具有群体性和持续性的特点,并且研究了算法中各参数间的影响和关系。由于现有同现模式挖掘方法在是否同现的判断上都采用了较武断的阈值,使得同现判断缺少对数据的容错性,概率建模是加强容错性的有效方法,如何直接对同现进行概率建模是下一步研究的方向。

参考文献

[1] Antunes C M, Oliveira A L. Temporal Data Mining: An Overview [C]//Proceedings of KDD Workshop on Temporal Data Mining. New York, USA: ACM Press, 2001:1-13.

[2] Miller H J, Han Jiawei. Geographic Data Mining and Knowledge Discovery [M]. Boca Raton, USA: CRC Press, 2009.

[3] 刘大有,陈慧灵,齐　红,等. 时空数据挖掘研究进展[J]. 计算机研究与发展,2013,50(2):225-239.

[4] Morimoto Y. Mining Frequent Neighboring Class Sets in Spatial Databases [C]//Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2001:353-358.

[5] Huang Yan, Xiong Hui, Shekhar S, et al. Mining Confident Co-location Rules Without a Support Threshold[C]//Proceedings of the 18th ACM Symposium on Applied Computing. New York, USA: ACM Press, 2003:497-501.

[6] Shekhar S, Huang Yan. Discovering Spatial Co-location Patterns: A Summary of Results [C]//Proceedings of the 7th International Symposium on Spatial and Temporal Databases. Berlin, Germany: Springer, 2001: 236-256.

[7] Estivill-Castro V, Lee I. Data Mining Techniques for Autonomous Exploration of Large Volumes of Geo-referenced Crime Data [C]//Proceedings of the 6th International Conference on Geocomputation. New York, USA: [s. n.], 2001:24-26.

[8] Estivill-Castrol V, Murray A T. Discovering Associations in Spatial Data——An Efficient Method Based Approach [C]//Proceedings of the 2nd Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin, Germany: Springer, 1998:110-121.

[9] Huang Yan, Zhang Pusheng. On the Relationships Between Clustering and Spatial Co-location Pattern Mining [J]. International Journal on Artificial Intelligence Tools, 2008, 17(1):55-70.

[10] Xiao Xiangye, Xie Xing, Luo Qiong, et al. Density Based Co-location Pattern Discovery [C]//Proceedings of the 16th International Conference on Advances in Geographic Information Systems. New York, USA: ACM Press, 2008:29-34.

[11] Worton B J. Kernel Methods for Estimating the Utilization Distribution in Home-range Studies [J]. Ecology, 1989, 70(1):164-168.

[12] Billiard F. Estimating the Home Range of An Animal: A Brownian Bridge Approach [D]. Chapel Hill, USA: University of North Carolina, 1991.

[13] Horne J S, Garton E O, Krone S M, et al. Analyzing Animal Movements Using Brownian Bridges [J]. Ecology, 2007, 88(9):2354-2363.

编辑　陆燕菲

(上接第 62 页)

[10] Bai Xiaopeng, Xue Nianwen. Building a Chinese Lexical Taxonomy[C]//Proceedings of the 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing. Tianjin, China: [s. n.], 2012.

[11] 何婷婷,张小鹏. 特定领域本体自动构造方法[J]. 计算机工程,2007,33(22):235-237.

[12] Wang Zuxing, Zhu Xiaoting, Lu Zhao. A Context-aware Automatic Chinese Transliterated Person Names Recognition Approach[C]//Proceedings of the 8th International Conference on Semantics, Knowledge and Grids. Beijing, China: [s. n.], 2012:143-149.

[13] Johnson M. PCFG Models of Linguistic Tree Representations[J]. Computational Linguistics, 1998, 24(4):613-632.

[14] Madsen R E, Kauchak D, Elkan C. Modeling Word Burstiness Using the Dirichlet Distribution [C]//Proceedings of the 22nd International Conference on Machine Learning. New York, USA: ACM Press, 2005:545-552.

[15] Dellschaft K, Staab S. Strategies for the Evaluation of Ontology Learning [C]//Proceedings of 2008 Conference on Ontology Learning and Population: Bridging the Gap Between Text and Knowledge. Amsterdam, Holland: IOS Press, 2008:253-272.

编辑　金胡考