第33卷 第22期 计 算 机 工 程 2007年11月
Vol.33 No.22 Computer Engineering November 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)22—0235—03 文献标识码：A 中图分类号：TP391.43

特定领域本体自动构造方法

何婷婷，张小鹏
(华中师范大学计算机科学系，武汉 430079)

摘 要：提出了一种自动构造特定领域本体的方法，该方法应用术语抽取和多重聚类技术。在术语抽取阶段，通过术语在专业语料与背景语料中出现概率的对比，采用LLR公式对术语进行评分，取得了更好的抽取效果。在层级关系发现过程中，采用上下文共现信息结合HowNet中词语的语义相似度，进行术语间相似度度量，力求获得术语间最合理的相关状况。同时改进了k-medoids聚类算法，更准确地发现术语的层级关系，进而构造出特定领域的本体。

关键词：本体；LLR；术语抽取；聚类；k-medoids

Approach to Automatical Construction of Domain Ontology

HE Ting-ting, ZHANG Xiao-peng
(Department of Computer Science, Huazhong Normal University, Wuhan 430079)

【Abstract】 This paper presents an approach to mining domain-dependent ontologies using term extraction and relationship discovery technology. There are two main innovations in the approach. One is extracting terms using log-likelihood ratio, which is based on the contrastive probability of term occurrence in domain corpus and background corpus. The other is fusing together information from multiple knowledge sources as evidences for discovering particular semantic relationships among terms. In the experiment, traditional k-mediods algorithm is improved for multi-level clustering. The approach to produce an ontology for the domain of computer science is applied and promising results are obtained.

【Key words】 ontology; LLR; term extraction; cluster; k-mediods

本体已经发展成为知识表示、知识管理、知识共享、知识复用的主流技术之一，正成为自然语言处理、Web信息检索、数据库和知识库的管理、异构数据集成、数字图书馆、GIS、语义Web等研究领域共同关心的一个核心。近几年来国内外学术界对此也越来越重视。由美国MIT、美国国防高级研究署(DARPA)和欧共体共同资助的W3C提出的语义网际网络研究计划中，领域本体结构及其相关技术是一个重要的研究领域。另外，从2001年起，欧共体又资助了一个称为OntoWeb的研究计划，主要是联合学术界和工业界共同研究和开发领域本体结构及其应用相关的技术[1～2]。

特定领域的本体能够形式化表达领域中的各种概念及概念之间的关系，从而将术语的语义表达出来，因而在语义查询和复杂推理方面发挥着重要的作用。比如，在机器推理和信息检索等系统中，特定领域的本体可以为这些系统提供该领域的术语、概念及其间的关系，从而提高整个系统的准确率。目前，本体的构造需要人工参与，是一项繁杂而费时的工作。因此迫切需要研究自动构造特定领域本体的策略和工具，至少可以为人工建造高质量的特定领域本体奠定良好的基础[3～4]。

本文提出了一种自动构造特定领域本体的方法：首先运用LLR评分规则抽取出领域术语，结合多种语义资源计算术语之间的相关度，然后通过多重聚类的方法得到术语之间的层级关系，进而构造出特定领域的本体。

1 相关工作

2001年，Govind和Chakravarthi等运用潜在语义索引(latent semantic index)方法进行本体的自动构造。他们采用潜在语义索引方法的目标是通过低维概念的空间代替高维的文档空间，从而获取术语与术语之间的关系。在Govind和Chakravarthi的方法中，概念被定义为一个语义类，即相关术语的集合。他们首先经过词频统计构造“术语-文档”矩阵，然后将“术语-文档”矩阵进行奇异值分解，分别得到术语矩阵U、奇异值矩阵S和文档矩阵V；最后通过术语矩阵U和文档矩阵V生成概念与术语之间的关系，从而构造出本体。该方法的特点是简洁，可以获得术语在文档中的共现关系，然而无法得到术语之间其他的明确关系。

Dekang Lin和Patrick Pantel提出了基于CBC(clustering by committee)聚类的领域概念发现方法。在该方法中，概念同样被定义为相关术语的集合。他们首先进行术语提取，统计术语共现频率，得到术语向量矩阵，进而运用CBC算法对术语进行聚类，获得领域概念。Lin和Pantel的方法旨在给出一种有效的聚类算法和概念自动发现的途径，从而获得术语之间的聚集情况，发现术语间的语义相关性，但是无法获取概念及术语间的关系描述，仅涉及到本体自动构造整个过程的一个层面，其思想具备很好的启发性。

2 本体自动构造

2.1 系统结构

图1大致描述了特定领域本体自动构造的过程：首先获取某一领域的相关文档，构成领域的语料资源，然后进行数据清理；在分词与词性标注的基础上，对词语进行评分，并

基金项目：国家自然科学基金资助项目(60442005)；教育部科学技术研究基金资助重点项目(105117)
作者简介：何婷婷(1964－)，女，博士、教授，主研方向：自然语言处理，数据库与数据挖掘；张小鹏，硕士研究生
收稿日期：2006-11-23 **E-mail：**zhangxiaopeng@mails.ccnu.edu.cn

根据词语得分确定领域的相关术语；利用统计信息，获得术语向量矩阵；通过多重聚类，获得术语间的层级关系，最终构造出特定领域的本体结构。

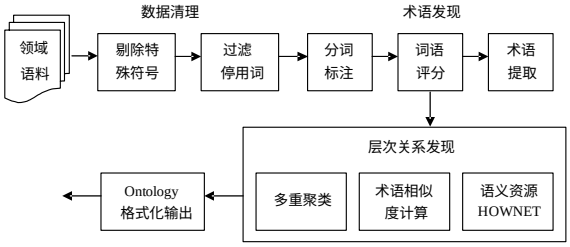

图 1　系统结构

2.2　术语抽取

本文的术语抽取方法基于以下认识：如果一个词语在特定领域的语料中出现的概率大于背景语料库中出现的概率，则该词是领域相关的。表 1 给出了计算机专业领域语料和背景语料中包含“算法”、“内存”、“路由器”、“数据库”等词的文档数。背景语料采用国家语委现代汉语语料库。

通过表 1 可以看出，“算法”、“内存”、“路由器”、“数据库”等词在计算机专业语料中出现的频率高于在背景语料库中出现的频率，则认为这些词具有较高的领域相关性。

表 1　文档频率统计

	算法	内存	路由器	数据库	文档总数
专业语料	89	78	73	69	200
背景语料	7	0	4	6	12 000
差　值	82	78	69	63	-11 800

术语抽取方法如下：(1)对语料进行预处理，并采用 SEGTAG 进行分词与标注；(2)统计词频，对词语进行领域相关度评分，根据评分决定领域术语。其中对词语的领域相关度评分，采用 LLR(log likelihood ratio)公式：

$$-2\log 2(Ho(p;k1,n1,k2,n2)/Ha(p1,p2;n1,k1,n2,k2))) \quad (1)$$

该公式用于度量备选假设 Ha 与零假设 Ho(假设每个术语在专业语料和背景语料中出现的概率相同)之间的差异度。这里 Ho 和 Ha 采用二项式模型。其中 $P=(k1+k2)/(n1+n2)$，$p1=k1/n1$，$p2=k2/n2$，$k1$ 表示术语在专业语料中出现的文档数，$k2$ 表示术语在背景语料库中出现的文档数，$n1$ 表示专业语料中文档总数，$n2$ 表示背景语料中文档总数。

2.3　术语层级关系发现

在获得领域术语的基础上，对术语进行聚类，获得术语间的层级关系。以术语上下文共现信息参考 HowNet 语义相似度作为术语之间相似度的度量，通过改进 k-medoids 算法对术语进行聚类，发现术语层级关系，过程如下：经过一次聚类，可以获得顶层的若干个类；对于得到的每个类，再次用该算法进行聚类，从而可以获得第 2 层的聚类情况；以次类推，可以获得多层的聚类分布状况，称之为多重聚类，具体计算中需要限定聚类重数。

2.3.1　术语相似度计算

根据 2.2 节提取的术语和专业语料统计共现词频，可以得到“术语-文档”矩阵 $\boldsymbol{M}[m][n]$(m 为术语个数，n 为专业语料中文档数目，$\boldsymbol{M}[i][j]$表示术语 i 在文档 j 中出现的次数)。该矩阵的每一行为一个术语向量

$$\boldsymbol{T}_i=(M[i][0],M[i][1],\cdots,M[i][k],\cdots,M[i][n]) \quad (2)$$

采用余弦相似度作为术语之间的相似度度量：

$$\cos(\boldsymbol{T}_i,\boldsymbol{T}_j)=\frac{\boldsymbol{T}_i\cdot\boldsymbol{T}_j}{|\boldsymbol{T}_i|\cdot|\boldsymbol{T}_j|} \quad (3)$$

表 2 给出了部分术语之间的余弦相似度。该方法考虑了术语的上下文共现信息，但可能损失部分术语之间的相似度。为了获得更合理的术语相似度，这里引入已有的语义资源中文知网中词语间的语义相似度进行补充。对 HowNet 中词语相似度计算，引用基于义项的计算方法。对于 HowNet 中未登录的词，设定它与其他词的相似度均为 0。最后，给出术语之间的相似度计算公式如下：

$$sim(\boldsymbol{T}_i,\boldsymbol{T}_j)=\frac{simA(\boldsymbol{T}_i,\boldsymbol{T}_j)+\alpha\cdot simB(\boldsymbol{T}_i,\boldsymbol{T}_j)}{2} \quad (4)$$

式中，$simA(\boldsymbol{T}_i,\boldsymbol{T}_j)$ 为术语之间余弦相似度；$simB(\boldsymbol{T}_i,\boldsymbol{T}_j)$ 为 HowNet 中术语之间的语义相似度；α 为可调参数，取经验值 α=0.5。表 3 给出了部分术语之间相似度的计算结果。

表 2　术语余弦相似度

术语 1	术语 2	相似度
缓冲区	磁盘	0.436 248
缓冲区	主存	0.579 771
内存	磁盘	0.434 059
服务器	网络	0.175 859

表 3　术语相似度

术语 1	术语 2	相似度
缓冲区	磁盘	0.510 913
缓冲区	主存	0.579 771
内存	磁盘	0.508 724
服务器	网络	0.318 716

2.3.2　聚类算法

在聚类过程中，对 k-medoids 算法进行了改进：在重新计算聚类中心时，不是简单地以距离每个类平均向量最近的术语作为类的新中心，而是首先根据类中元素的分布，决定该类中聚集度最大的 p 个术语，然后将距离这 p 个术语的平均向量最近的术语作为该类的新中心。这样可以使得聚类过程更快地收敛，聚类结果更接近术语的原始聚集分布状况。算法具体描述如下：

(1)选择 m 个术语作为初始聚类中心 $C1,C2,C3,\cdots,Cm$。

(2)对于每个术语 $\boldsymbol{T}_j$，运用式(4)计算其与每个聚类中心的相似度，将该术语加入与其相似度最大的类中。

(3)按如下方法重新计算 m 个类的中心：

1)计算类 i 中所有术语(i=1,2,⋯,m)的平均相似度 $avgsim[i]=\frac{1}{n}\sum_{k=1}^{n}sim(T_k,C_i)$($n$ 为类 i 中术语个数)。将 m 个平均相似度的最大值作为最大平均相似度 max_$avgsim$。

2)选出类 i(i=1,2,⋯,m)中与类 i 中心相似度最大的 p 个术语，p 由下式决定：

$$p=m*\frac{avgsim[i]}{\max\_\ avgsim}$$

3)计算 p 个术语的平均向量，将与其最近的术语作为类 i 的新中心。

(4)如果聚类中心未趋于稳定，即新的聚类中心与上一次聚类中心差距大于某阈值，则转到(2)。

(5)得到 m 个类及其中心向量，算法结束。

3　实验分析及评价

本文采用的专业语料由 200 篇(3.12MB)计算机期刊的论文组成，背景语料库采用国家语委现代汉语语料库(1996~2001)，共 12 000 篇文章(33.5MB)。

3.1　术语抽取

实验中，对术语评分规则 LLR 和 TF.IDF 进行了比较。

表 4 给出了部分术语的评分情况，结果采用相对百分比。表中粗体词语表示笔者判断得到的计算机专业领域领域相关度高的术语。从中可以看到，LLR 方法优于 TF.IDF 方法。

表 4　计算机领域术语评分

术语	领域 DF	背景 DF	LLR	TF.IDF
路由	73	4	98.6	66.8
总线	56	1	99.9	70.4
缓存	36	2	94.6	53.5
请求	17	188	67.5	72.6
线路	31	232	56.2	65.7

笔者通过人工方法对结果进行评估。首先从专业语料中构造出一张包含 286 个术语的术语表，然后采用召回率(R)、准确率(P)和 F-度量(F)对结果进行评测。

图 2 给出了召回率(R)、准确率(P)和 F-度量(F)随 LLR 阈值变化的情况。

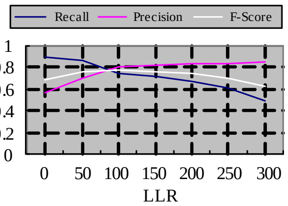

图 2　召回率、准确率和 F-度量变化情况

3.2　层级关系发现

根据图 2，选择 F-score 的阈值为 0.76，获得 216 个领域相关的术语。聚类过程中，限定聚类重数为 2。每个类以该类中出现频率最高的词命名，下面给出了部分术语之间的层级关系。

<-网络->

::路由器 路径 主干网 速率 中继 广域网 延迟 分组 交换 局域网 交换网 接入网 策略 关键词 网格 *网络* 拓扑 骨干网

::*搜索* 引擎 网页 脚本 主机 万维网 网站 网址 带宽 阻塞 端口 监听 信息网

::防火墙 流量 计算中心 域名 主页 *服务器* 拷贝 电子邮件 邮件 邮箱 日志 数据流 语音 共享

::智能性 *信息港* 因特网 智能化 电话网 宽带 调制 解调 频率段 数字网<-计算机->

::体系 结构 总线 硬件 接口 中断 控制器 寄存器 运算 时间段 数据 共享 实时 系统 调度者 *进程* 子系统 分系统 硬盘

::*计算机* 芯片 大型机 巨型 磁盘 软盘 光盘 主存 吞吐 软驱 外存 内存 存储器 存储 缓冲 缓存 输出 驱动 驱动器 适配器 网卡 交换机

<-程序->

::矩阵 数组 指针 字符串 *参数* 静态 访问 编译器 变量 字符标量 流程图 流程 结构图 逻辑图 运算符

::控制台 *编程* 源程序 程序包 子程序 程序性 源代码 向量 编译 调试 调试器 程序员 瓶颈 数据包

::编程者 查询 分析器 数据库 数据表 冗余 备份 标识符 计数器 初始化 *程序* 语句 表达式 操作数

::*软件* 软件包 构建 复用 重构 组件 构件 重用性 模块 封装性 封装 继承 私有 实例 对象

通过人工评估，第 1 级聚类的准确率达到了 76.7%，第 2 级聚类的平均准确率达到了 70.3%。

4　结束语

本文提出了一种特定领域本体的自动构造方法。该方法运用了术语抽取和聚类技术，在术语抽取阶段，采用 LLR 公式对词语进行评分，提高了抽取的准确率。在层级关系发现过程中，以术语在上下文中的共现信息结合 HowNet 中词语的语义相似度为辅助，进行术语相似度计算，力求获得术语间最合理的相关状况。

今后的工作中，还有以下几点值得思考：(1)减小对语料处理的粒度；(2)引入启发式规则，更准确地发现术语以及术语之间的关系；(3)发现并标注多种关系。

参考文献

1. Lin D, Pantel P. Induction of Semantic Classes from Natural Language Text[C]//Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA. 2001: 317-322.

2 Maedche A, Staab S. Mining Ontology from Text[C]//Proc. of the 12th International Workshop on Knowledge Engineering and Knowledge Management. 2000.

3 Mani I. Automatically Inducing Ontologies from Corporate[C]// Proceedings of the 3rd International Workshop on Computational Terminology, Geneva. 2004.

4 Srivastava S, Lamadrid J G. Extracting an Ontology from a Document Using Singular Value Decomposition[R]. Association of Computer and Information Science and Engineering Departments at Minority Institutions, 2001.

~~~~~~

(上接第 231 页)

参考文献

1 Wang Y. The OAR Model for Knowledge Representation[C]//Proc. of IEEE Canadian Conference on Electrical and Computer Engineering, Ottawa, Canada. 2006: 1696-1699.

2 Zhao Ke, Hu Gangwei, Xu Wei, et al. Research on Concept and Concept Relation Model Used for Semantic Disambiguation of Natural Language[C]//Proceedings of the 5th IEEE International Conference on Cognitive Informatics, Beijing. 2006: 320-331.

3 钱　军. 句式意义——句法与语义关系的若干理论问题[J]. 外语研究, 2004, (2): 5.

4 郑远汉. 省略句的性质及其规范问题[J]. 语言文字应用, 1998, (2): 10-11.
~~~~~~