一种基于语料库和互信息的本体学习方法*

李向阳
(华侨大学 工商学院,福建 泉州 362021)

摘 要:提出了一种应用中文自由文本作为知识源的本体构造方法,将采用该方法分词后得到的词汇分别计算,进而得到在样本文本和日常语料库中的出现概率估计值,将二者对比得到出现频率的显著性指标,由此自动识别并提取领域用词汇,再应用互信息分析识别领域词汇之间的结合特性。它可自动建立可能的领域本体词汇及词汇之间基本关系的集合,同时还可构造出基于领域词汇和它们之间结合度的领域词图,为进一步进行人工本体构造提供方便的可视化界面。该成果可为实现大规模基于内容的知识管理提供自动化/半自动化本体支持。

关键词:领域本体;本体自学习;互信息
中图分类号:TP182 **文献标识码**:A **文章编号**:1674-7720(2010)10-0074-04

A corpus and mutual information based ontology learning method

LI Xiang Yang
(College of Business Administration, Huaqiao University, Quanzhou 362021, China)

Abstract: A computer aided domain ontology building method from free Chinese domain text is developed. The key steps are exemplified by a group of experimental data. The domain word elements are identified semi-automatically through comparing the appearing probability of them in sample text with which in the corpus. Then the domain words composition ability among the identified word elements been calculated using mutual information. Domain ontology words and the relationship among them can be generated through using the data produced from the steps. A directed weighted word graph can also been generated, which could be used as a good visual ontology draft for the ontology builders works on. The result could be used to support large scale content based knowledge management.

Key words: domain ontology; ontology learning; mutual information

领域本体可为实现基于内容的知识管理以及基于语义的全文检索提供语义支持。本体(ontology)是领域知识的概念化显式说明[1],目前已被公认为实现知识表示、知识共享和知识交换的基础,在知识管理、智能检索、语义 Web 及专家系统等领域发挥着重要作用。领域本体实际上是领域知识的语义模型,包含描述知识的重要术语词汇,同时定义了它们之间的语义关系。由领域专家手工构造本体存在耗时多、成本高、难以适应快速的知识更新要求等诸多问题;另一方面,人工构造的本体还具有较大的主观性,规范性、统一性较差,不便于与其他相关本体的集成。

国外基于英文和其他语种的本体自学习已有大量研究[2-3],但基于中文的本体自学习还不多。参考文献[4]提出了一种基于句式规则的自举本体学习。而本文提出的方法应用中文语料库,通过统计方法识别本体词汇及它们之间的可能关系,构造出了一个本体的初始框架。

1 基本原理与系统结构

本文提出的方法基于以下 4 个观察和假设:(1)领域文档中相关的领域词汇频繁出现;(2)高频同现的术语具有语义联系;(3)专业术语的组词规则中包含可用的语义信息;(4)修改一个初具规模的语义模型的草案比从头建立一个语义模型要容易得多。领域专业教材及相关文档资料是人们认知和学习一个全新领域的重要知识资源,它同样可以作为计算机自动获取本体的知识资源。基于

* 基金项目:华侨大学科研基金资助项目(04HZR18)

这些原则自动地获取领域词汇并发现它们之间的组合关系，不但可直接用于对语义要求不太严格的知识管理中，而且也可为专家构造更精确的本体提供良好基础。

本实验完成的计算机辅助本体构造过程分为三个基本步骤：选择和准备知识资源、提取领域术语词汇、建立词汇术语之间语义关系。第一步主要由操作员人工操作实现，它不需要深入的领域专业知识。选择的标准是高频集中出现领域词汇的文档资料，如专业教材目录、术语索引表等可作为初始学习的知识源。电子课本、培训资料等都可以作为选择领域的对象，其主要目的是将各种不同格式的文档资料转换为统一的 text 格式，作为系统输入的知识源数据；第二步，系统先应用自然语言语料库自动识别领域词汇，再由专家进行确认或修改；第三步，系统先分析所提取的词汇在本体中的可能角色，建立领域词汇关系图，再由领域专家用户进行确认和修改。最终得到一个领域本体的基本框架。

以下两节详细说明第二步和第三步的算法。说明中引用参考文献[5]中第一章中 1.3.4 小节“数据库系统的分类”作为样本(后文简称为样本文档)。

2 用词频特征值识别和提取领域词汇

一般文本挖掘方法在识别词汇时，事先筛选某些常用词汇作为高频词，它们在识别过程中被排除。这里不采用这种方法，因为中文的领域词汇通常也会使用某些常用字/词，并对它们赋予新的领域含义。而采用词汇在样本文本与普通文本出现的概率对比，当比值显著高于语料库中的频率时，推断它为领域词汇。

2.1 据语料库识别词汇出现的正常概率

本实验应用北大-富士通研制的《人民日报语料库》1998 年 1 月的带词性标注的语料库[6]，它共包含 948 400 词(本实验统计结果)。以它作为常用词库的参照，统计出其中各词汇的出现频率，形成中文自然语言词汇的一般使用概率估计值，其计算公式为：

$$\bar{P}_{W_i}=C_{W_i}/C_{total} \quad (1)$$

其中 $\bar{P}_{W_i}$ 为目标词汇 W_i 在语料库中出现的概率估计值，C_{W_i} 为词在语料库中出现的总次数，C_{total} 为语料库中总词汇量。本实验应用编写的程序计算出语料库中所有词汇的一般使用概率估计值。

2.2 样本中词汇出现的概率与特征值

本实验对样本内容应用 ICTCLAS 分词系统[7]进行分词，计算总词汇量和各个词的出现资料数，然后按照下文 3.1 中描述的相同的方式求出每个词在该样本中出现的概率估计值 SP_{Wi}。词汇在样本文档中出现的概率估计值与一般使用概率进行对照，以前者除以后者得到值作为显著性指标，得到领域词汇特征值 E_{Wi}，其计算公式表示为：

$$E_{Wi}=SP_{Wi}/\bar{P}_{W_i} \quad (2)$$

表 1 中按照由大到小顺序列出前 21 个词汇的样本概率、语料库概念和特征值。

表 1 实验样本文档的特征值

样本词	样本概率/SP	语料库概率/P_{wi}	特征值/E
并行	0.020 833	0.000 002 1	9 920.476
分布式	0.006 944	0.000 001	6 944
浏览器	0.005 208	1.05E-06	4 960
服务器	0.026 041	6.32E-06	4 120.411
数据库	0.045 138	1.265E-05	3 568.221
DBMs	0.005 208	0.000 001	5 208
扩展性	0.003 472	0.000 001	3 472
集中	0.003 472	0.000 232	14.968 1
良好	0.003 472	0.000 284 7	12.195 72
或	0.006 944	0.000 708 6	9.800 158
技术	0.008 68	0.000 989	8.776 276
个	0.006 944	0.002 776 3	2.501 216
的	0.065 972	0.057 451 5	1.148 308
与	0.003 472	0.003 062	1.133 903
在	0.013 888	0.012 678 2	1.095 425
是	0.010 416	0.010 382 7	1.003 203
不	0.003 472	0.004 778 6	0.726 577
有	0.003 472	0.004 893 5	0.709 513
为	0.003 472	0.004 998 9	0.694 547
一	0.003 472	0.007 734 1	0.448 923
了	0.005 208	0.012 185 8	0.427 383

在实验中考虑到两个方面：一是领域知识源的样本词汇量比语料库词汇量小得多，为减少因样本空间带来的误差，只对样本文档中出现频率次数两次以上的词进行分析；二是有些词汇，如“分布式”、“DBMS”、“扩展性”等，在平均概率估计样本中不存在，为防止除数为零且便于计算，统一取其平均概率估计值为百万分之一。

从表 1 可以看出，“的”、“个”、“与”、“在”、“是”等平均概率高的词汇在领域词汇中出现的概率也高，但其概率与平均概率相比差距不大；而“并行”、“服务器”、“数据库”等词汇在测试文档中以高概率估计值出现，与平均概率相比存在几千倍的差值，因此它们显著地具有领域词汇的特征。

2.3 识别领域词汇特征值阈值

计算得到各词汇的样本概率和特征值后，特征值越大的词越可能成为领域词汇候选词，若特征值选取过大，则可能漏掉一些领域词汇，减少了招回率；当特征值选取过小时，则非专业词也可能被选中，有损于系统的准确率(P)。系统抽取质量的评判指标是二者的综合，称为 F 度量：

$$F=(\beta^2+1)\times P\times R/((\beta^2\times P)+R) \quad (3)$$

其中，β 为对精度的偏重量，一般取 $\beta=1$。

阈值的选取以 F 值达到最大作为差别的依据，它是一个经验值，与文档知识源相关。本实验经过多次试验证明，F 取 13 作为阈值时能够在招回率与准确度之间达到较好的折中。

3 组合领域词汇的识别

从构词法角度来看，专业领域中的术语有三种基本构成形式：一是给普通词汇赋予新的领域含义；二是创建一个全新的词；三是以前两种形式为词干加上前缀或后缀形成新词。

第一种领域词汇通过分词系统自动划分为一个独立的词，在语料库中也会出现，它可以通过上一节的词频对比分析识别得到。第二种领域词汇在自动分词系统中无法分出，在语料库中也没有该词，它由若干单字或常用词组合而成。第二种和第三种可应用信息论中的互信息，自动地从样本文档中识别。

3.1 用互信息识别高频词汇组合

信息论中互信息反映了一种信息与另一种信息相关联的程度。可用式(4)表示为：

$$M(a,b)=\log_2(P(a|b)/p(a)) \quad (4)$$

其中 $P(a)$、$P(b)$ 分别表示事件 a 和 b 出现的概率，$P(a|b)$ 为事件 a 相对于事件 b 的条件概率。在本系统中，以样本文档中总词数 cntTotal 为基数，以词出现的次数 c 除以总词数做为概率估计值。$P(a|b)$ 用 a 与 b 同现次数除以 b 出现次数作为估计值。

本实验对文档先后同现两次以上的词进行互信息统计分析。前后同现的两个词分别记为 w_1 和 w_2，应用互信息计算公式通过程序统计分析得到表 2。

表 2 词汇同现互信息分析

w_1	w_2	C_{W12}	C_{w1}	C_{w2}	$M(w1,w2)$
集	中式	2	2	2	8.170
结	点	3	3	3	7.585
集中	式	2	2	3	7.585
三	层	3	3	6	6.585
多	处理机	2	8	2	6.170
微型	计算机	2	2	8	6.170
客户	机	7	9	7	6.0
机	/	3	7	4	5.948
组	数据	2	2	10	5.848
事务	处理	2	3	8	5.585
并行	计算	2	12	2	5.585
应用	程序	3	12	3	5.585
层	结构	4	6	9	5.415
/	服务器	3	4	15	4.848
提高	系统	3	4	23	4.231
分布式	数据库	3	4	26	4.055
式	数据库	2	3	26	3.885
数据库	系统	13	26	23	3.646
并行	处理	2	12	8	3.585
并行	计算机	2	12	8	3.585
应用	服务器	3	12	15	3.263
数据库	服务器	4	26	15	2.562
并行	数据库	3	12	26	2.470
应用	系统	2	12	23	2.061

3.2 构造高频词汇组合关系图

表 2 中要排除 w_1 和 w_2 都不在表 1 中出现的项。以表 2 中相关的词汇两两组合，以它们之间的组合关系为边，以相关度为权值构造一个有向加权多图。图 1 所示为部分高频词及其可能组合。

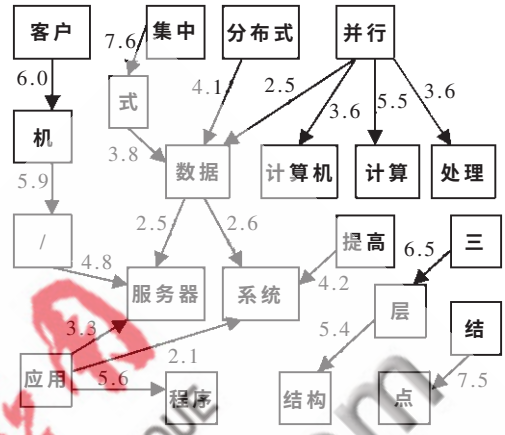

图 1 词汇高频组合关系图

图 1 有两种意义：一是作为领域专家选取有效领域词汇，去除无效组合的图形化用户界面。如图中的“客户”与“机”、“集中”与“式”、“结”与“点”等有高二词相关度，有较大的概率组合成词“客户机”；“三”、“层”、“结构”和“客户”、“机”、“/”、“服务器”等具有高路径相关度，在图中的整个路径具有较大的成词概率，分别可能构成“三层结构”和“客户机/服务器”，它们可在界面中标记确认；而“提高”与“系统”虽然也有较高的二词相关度，但它们不构成一个领域词，可通过该界面去除。第二种意义是，经过用户的领域词汇确认操作后，可以得到一个本体的基本可视化框架，可以对这些词汇和它们之间的关系添加其他语义信息。图 2 是基于本文算法自动取得领域词汇后，对它们之间的语义关系作进一步确定和编辑的用户界面。

本文通过实验例示了基于中文自由文本提取领域

图 2 专业词汇关系构造器用户界面

词汇并构造领域本体基础框架的关键过程。实验证明该方法能在可接受的运算时间内有效地识别本体词汇并建立本体草案，它可直接应用于对本体要求不太严格的基于内容的知识管理中，也可应用于软件需要分析的早期阶段。本文介绍的成果还有许多值得进一步研究和开发之处，如词汇高频组合关系图还可进一步实现自动生成，还可发掘同现词汇之间更深层的语义关系等。

参考文献

[1] MAEDCHE A, STAAB S. Ontology learning for the semantic web[J]. IEEE Intelligent Systems, 2001, 16(2): 72-79.

[2] NAVIGLI R, VELARDI P, GANGEMI A. Ontology learning and its application to automated terminology translation[J]. IEEE Intelligent Systems, 2003, 18(1): 22-31.

[3] SHAMSFARD M, BARFOROUSH A A. Learning ontologies from natural language texts[J]. International Journal of Human Computer Studies, 2004, 60(1): 17-63.

[4] 贾秀玲，文敦伟. 面向文本的本体学习研究概述[J]. 计算机科学，2007，34(2)：181-185.

[5] 王珊，李盛恩. 数据库基础与应用[M]. 北京：人民邮电出版社，2002：11-12.

[6] 人民日报语料库[EB/OL]. [2001-05-10]. http://www.icl.pku.edu.cn/icl_groups/corpus/dwldform1.asp.

[7] 张华平，刘群. 计算所汉语词法分析系统 ICTCLAS. http://sewm.pku.edu.cn/QA/reference/ICTCLAS/FreeICTCLAS/, 2002.

(收稿日期：2010-01-27)

作者简介：

李向阳，男，1971年生，在读博士生，副教授，主要研究方向：知识处理与知识工程、语义 Web。