一种改进的领域本体分类算法

李 刚 钱省三 叶春明
(上海理工大学管理学院 上海 200093)

摘 要 本体学习技术的研究目前还处于探索阶段。研究了知识生产领域中本体学习技术的应用,提出了本体学习中领域本体的分类算法。本算法同时考虑了概念之间的语义相似度与结构相关度,并以"分类量化值"作为领域本体的分类标准。实验证明,本算法较之其它相关算法更为有效。

关键词 本体学习技术,领域本体,分类算法
中图法分类号 C931.6 **文献标识码** A

Improved Classification Algorithm of Domain Ontology

LI Gang QIAN Xing-san YE Chun-ming
(Management Department, Shanghai University for Science and Technology, Shanghai 200093, China)

Abstract Ontology learning technology is still in the exploratory stage. The paper researched the ontology learning techniques in the field of knowledge production, and proposed a classification algorithm of ontology. As the algorithm taking into account both the semantic similarity and the structure similarity of the concept, using the "quantitative classification of value" as the classification standards, experiment shows that the algorithm is more effective than other algorithms.

Keywords Ontology learning technology, Domain ontologies, Classification algorithm

知识经济是以知识为基础的经济,直接依赖于知识的生产、扩散与应用。与传统工业生产依赖于资金与自然资源的投入不同,知识经济时代经济的发展越来越依赖于知识资源的占有与配置。在传统经济向知识经济转变的过程中,知识及其生产者已经成为核心生产要素。

曼纽尔·卡斯特[1]认为,在新的知识经济时代,生产力的来源在于产生知识、信息处理与象征沟通的技术。针对知识本身的知识生产活动,就是生产力的主要来源。

侯象洋认为[2],知识生产管理指的是通过组织化的知识生产手段,利用知识生产的设施、技术、工具、人力(脑力),为了将原始的知识、数据等知识生产的"原料"稳定、高效地转化为能够解决客户问题、能够为客户所接受并为此付费的知识产品,而建立起来的管理体系以及以通过这个管理体系对知识生产过程所进行的各项管理活动。

近年来,人工智能中本体应用的兴起,为知识生产管理提供了有利的武器,尤其在领域本体的构建与分类管理方面。

本体的应用是建立在领域本体构建基础之上的,但本体的构建却是繁琐而复杂的工作。虽然现在本体构建工具日趋成熟,从最早的 Ontoligua 到 protege2000, Ontoedit,但是这些工具支持的仍然是手工构建本体的方式。如何进行领域本体的构建及分类,引起了众多学者的关注,本文利用本体学习技术对领域本体构建过程中的分类算法进行探讨。

本体的构建过程纷繁复杂,并且需要耗费大量的人力和物力,为了利用知识获取技术来降低本体构建的开销,近年来,利用本体学习技术进行本体的自动化建模逐渐成为计算机科学领域的一个研究热点。本体学习(Ontology learning)技术[3]是综合本体工程技术、机器学习技术和统计等技术自动或半自动地构建本体。本体学习涉及到从输入数据中提取本体学习内容(概念知识)并用这些内容构建本体。目前国内这方面的研究文献较少,国外则有较多文献[4-6]涉及此领域的研究,如 Maedche[7]等人介绍了一种文本自动抽取为本体的工具环境,提出了本体获取的框架,Missikoff[8]等人提出了本体学习工程中的集成方法,即从一组文本集中抽取出领域相关术语,再使用通用本体 WordNet 中的概念对其解释,确定术语之间语义关系。

目前关于领域本体构建与分类的研究虽然很多,但本体的半自动化构建技术仍然很不成熟,需要进一步研究与改进。基于此,本文提出一种改进的领域本体分类的算法,以期对本体学习技术在知识生产领域的应用进行探索。

1 领域本体分类算法

领域本体的分类是本体半自动化建模的关键步骤。由于领域本体中的概念及术语是抽象的字符,如果仅仅人为地进行分类不仅浪费大量的时间,而且对术语的归类也不够准确。

为了准确地对概念进行分类,本文采用术语量化技术,从语义以及结构这两个角度对术语之间的紧密程度进行量化。

到稿日期:2009-04-02 返修日期:2009-07-25 本文受上海市教委科研创新基金项目(08YS103),上海市重点学科项目(S30504)资助。
李 刚 博士后,研究方向为知识生产管理、企业信息化;钱省三 教授,博士生导师,研究方向为工业工程、知识生产管理;叶春明 教授,博士生导师,研究方向为工业工程、企业信息化。

用“分类量化值”表示两个概念的紧密程度，“分类量化值”由两个指标决定，一个是两个概念之间语义上的相似程度，即语义相似度；另一个是两个概念之间结构上的相关程度，即结构相关度。求得“分类量化值”后，对领域本体进行分类。

定义 1 概念 $A=\{a_1,\cdots,a_m\}$ 是 m 个术语的集合；$B=\{b_1,\cdots,b_n\}$ 是 n 个术语的集合，a_i 是 A 中的一个术语，b_j 是 B 中的一个术语。

定义 2 概念的分类量化值记作 $S_{catogory}$，分类量化值由概念语义相似度（记作 $S_{semantic}$）和结构相似度（记作 $S_{structure}$）表示。

定义 3 术语间语义相似度的计算，采用基于 Wu-Palmer[9] 的语义相似度算法。对于两个术语 a_i,b_j，其语义相似度（$S_{semantic}$）计算公式为

$$S_{semantic}(a_i,b_j)=\frac{2\times depth(lso(a_i,b_j))}{depth(a_i)+depth(b_j)} \tag{1}$$

其中，$lso(a_i,b_j)$ 是术语 a_i,b_j 的共同祖先概念，$depth(a_i)$ 和 $depth(b_j)$ 分别表示术语 a_i 和 b_j 在词典语义树中的深度。

定义 4 术语间结构相关度的计算

$$S_{structure}(a_i,b_j)=\sum_{k=1}^{r}p_k n_k(a_i,b_j) \tag{2}$$

其中，r 为 a_i,b_j 间的联系数，这里的联系主要包括关联、继承和各种依赖等。p_k 是第 k 种联系的强度（如可设 $p_{关联}=0.7$，$p_{继承}=1$……等），$n_k(a_i,b_j)$ 是 a_i,b_j 之间的第 k 种联系数量。

定义 5 概念 A,B 之间的分类量化值记作 $S_{catogory}$

$$S_{catogory}(A,B)=\frac{\alpha\sum_{i=1}^{m}\sum_{j=1}^{n}S_{semantic}(a_i,b_j)+\beta\sum_{i=1}^{m}\sum_{j=1}^{n}S_{structrue}(a_i,b_j)}{m\times n} \tag{3}$$

其中，α,β 值分别是语义相似度系数，结构相关度系数，$\alpha+\beta=1$ 且 $\alpha,\beta\in(0,1)$，一般情况下设 $\alpha=\beta=0.5$。

可以设定一个阈值 ω，当 $S_{catogory}(A,B)>\omega$ 时，表示两个概念 A,B 属于同一分类。

对于多个概念的分类算法，可以在每对领域本体之间运行以上算法，将求得的分类量化值构成矩阵，根据矩阵中分类量化值的大小对多个概念进行归类。

2 算法实例

这里以两个概念 A=“离心泵”，B=“离心机”为例，其共同祖先概念为 $lso(a_i,b_j)$=“旋转式流体机械”。3 个概念在本体库中深度分别是：$depth(A)=8$，$depth(B)=9$，$depth(los(A,B))=6$，仅采用传统的 Wu-Palmer 语义相似度计算，求得 A,B 两概念的语义相似度为 0.70。

结合 A,B 的概念描述：

离心泵是利用离心力，高速旋转的叶轮叶片带动流体转动，将流体甩出，从而达到输送目的的机械装备。将 A 定义为 4 个术语的集合，即 A=｛离心力，叶轮，流体，甩出｝。

离心机是利用离心力，分离液体与固体颗粒或液体与液体的混合物中各组分的机械。将 B 定义为 3 个术语的集合｛离心力，液体，分离｝。

分别利用本文提出的启发式算法的式(1)和式(2)进行语义相似度计算、结构相关度计算，分别求得 A,B 两概念的语义相似度为

$$S_{\substack{semantic\,a_i b_j\\ i=1\text{ to }4,j=1\text{ to }3}}=\begin{bmatrix}S_{semantic}a_1b_1 & S_{semantic}a_1b_2 & S_{semantic}a_1b_3\\ S_{semantic}a_2b_1 & S_{semantic}a_2b_2 & S_{semantic}a_2b_3\\ S_{semantic}a_3b_1 & S_{semantic}a_3b_2 & S_{semantic}a_3b_3\\ S_{semantic}a_4b_1 & S_{semantic}a_4b_2 & S_{semantic}a_4b_3\end{bmatrix}=\begin{bmatrix}1 & 0.5 & 0.6\\ 0.4 & 0.25 & 0.4\\ 0.6 & 0.8 & 0.6\\ 0.7 & 0.6 & 0.9\end{bmatrix}$$

结构相似度为

$$S_{\substack{structure\,a_i b_j\\ i=1\text{ to }4,j=1\text{ to }3}}=\begin{bmatrix}S_{structure}a_1b_1 & S_{structure}a_1b_2 & S_{structure}a_1b_3\\ S_{structure}a_2b_1 & S_{structure}a_2b_2 & S_{structure}a_2b_3\\ S_{structure}a_3b_1 & S_{structure}a_3b_2 & S_{structure}a_3b_3\\ S_{structure}a_4b_1 & S_{structure}a_3b_4 & S_{structure}a_4b_3\end{bmatrix}=\begin{bmatrix}1 & 0.68 & 0.57\\ 0.25 & 0.14 & 0.12\\ 0.62 & 0.7 & 0.46\\ 0.83 & 0.75 & 0.87\end{bmatrix}$$

这里设语义相似度系数与结构相关度系数为 0.5，即 $\alpha=\beta=0.5$，利用式(3)求得

$$S_{catogory}(A,B)=\frac{0.5\times 7.35+0.5\times 6.99}{4\times 3}=0.5975$$

A,B 概念的语义相似度是 0.5975。

由以上计算过程与计算结果可以看到，本文提出的启发式算法要优于现有的单纯基于 Wu-Palmer 的语义相似度算法，本文由于考虑了构成概念“离心泵”和“离心机”的所有术语的语义相似性与结构相关性，因此求得的算法更为精确，“离心泵”和“离心机”虽然都是利用离心力原理的机械设备，但前者是为了将流体甩出，而后者是为了将固液或液液分离，显然概念“离心泵”和“离心机”之间的匹配较差，本节算法求得的结果更为精确和有效。

结束语 本文提出的领域本体分类算法，较之文献[10-13]的研究方法，不仅考虑了概念之间语义相似度，还考虑了概念之间结构上的相关程度即结构相关度，通过量化求得“分类量化值”，更有利于将领域本体量化，提供领域本体的分类依据。并通过实验验证了分类算法的有效性。

参 考 文 献

[1] 曼纽尔·卡斯特. 网络社会的崛起[M]. 北京：社会科学文献出版社，2001

[2] 侯象洋. 知识生产管理：整体包装解决方案(CPS)的大规模定制[J]. 包装世界，2008(4)

[3] Du X Y, Li M, Wang S. A survey on ontology learning research [J]. Jourrml of Software, 2006, 17(9): 1837-1847

[4] Pantel P, De Kang-lin. A statistical corpus-based term extractor [C]// Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence. 2001: 36-46

[5] Buitelaar P, et al. A Protege Plug in for ontology extraction from text based on linguistic analysis[C]// Proceedings of the 1st European Semantic Web Symposium. 2004

[6] De Kan-lin, Pantel P. Concept discovery from text [C] // Proceedings of the 19th International Conference on Computational Linguistics. 2002: 1-7

[7] Maedche A, Volz R. The Text-To-Onto ontology extraction and

maintenance environment[C]//Proceedings of the ICDM Workshop on Integrating Data Mining and KnowledgeManagement. California,2001

[8] Missikoff M, Navigli R, Velardi P. Integrated approach to web ontology learning and engineering[J]. IEEE Computer,2002,35(11):60-63

[9] Budanitsky A, Graeme H. Evaluating WordNet-based measures of semantic distance[J]. Computational Linguistics, 2006, 32(1):13-47

[10] Cimiano P, Hotho A, Staab S. Learning concept hierarchies from text corpora using formal concept analysis[J]. Journal of Artificial Intelligence Research,2005(24):305-339

[11] Bisson G, Nedellec C, Canamero D. Designing clustering methods for ontology building: The Mo'K workbench[C]// Proceedings of the ECAI 2000 Workshop on Ontology Learning(OL'2000). 2000

[12] Maedche A, Staab S. Ontology learning[C]// Proceedings of 14th European Conference on Artificial Intelligence. 2000

[13] Faure D, Nedellec C. A corpus-based conceptual clustering method for verb frames and ontology acquisition[C]//Proc. LREC-98 Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications, European Language Resources Distribution Agency. Paris,1998

[14] 孙亮,任小康. 基于本体的图像语义检索模型[J]. 重庆工学院学报:自然科学版,2009,23(1):127-131

(上接第 224 页)

2.6×10^{-6},即令 $p_0=2.6\times10^{-6}$,若这时令可信度 δ 分别取 0.93,0.95,0.97,0.99,则按式(20)计算得到表 1 所列出的结果。

表 1 基于字的 Unigram 模型训练样本规模

δ	0.93	0.95	0.97	0.99
$\Phi^{-1}(\frac{1+\delta}{2})$	1.81	1.96	2.17	2.58
N_w	1.13×10^7	1.33×10^7	1.63×10^7	2.30×10^7

表 1 表明,当训练语料规模达到 1130 万时,以式(3)所表示的 Unigram 模型,对所描述语言的估计有 93%的把握能够达到误差($\varepsilon=1.3\times10^{-6}$)要求。

如果要建立基于词的 Unigram 模型,则由于词的数量巨大,低频词的频率会更低,根据对 1995 年 2—7 月和 12 月的约 1600 万的《人民日报》语料的统计,二字以上的词出现频次在 15 次以上的词共有 21592 个;出现频次在 2～14 次的词共有 46637 个,这其中包括了不少的常用人名、地名以及数字;出现 1 次词共 9327 个,其中主要是一些人名和地名以及数字,当然也有一些和社会发展相适应的新词开始出现,比如"黑客"、"劝退"等。因此可以看出,低频词还是占据大多数,若 14 次以下的词都称作低频词,它们的频率约为 8.75×10^{-7},取 $p_0=8.75\times10^{-7}$,若仍然取可信度 $\delta=0.93,0.95,0.97,0.99$,则按式(20)计算得到表 2 所列出的结果。

表 2 基于词的 Unigram 模型训练样本规模

δ	0.93	0.95	0.97	0.99
$\Phi^{-1}(\frac{1+\delta}{2})$	1.81	1.96	2.17	2.58
N_w	3.38×10^7	3.96×10^7	4.84×10^7	6.84×10^7

若将 p_0 值取得更小,则要求的语料规模会进一步加大。由于表 2 考虑的建模语言单位是词,而根据对《现汉》的统计[8],二字词占总词数的 66.9%,一字词占 14.7%,三字词占 9.2%,四字词占 8.4%,五字以上的词占 0.8%。所以,其平均词长为 2.15,即使三字以上的长词出现较多,平均词长估计也不会超过 3。由此可见,建立以词为单位的 Unigram 模型所需训练语料规模最低估计为 $N=2.15\times N_w$。

如果考虑将式(6)中的绝对误差改为相对误差,即对其中的不等式 $\left|\frac{X}{N_w}-p\right|<\varepsilon$ 两边各除以 p,则式(6)变为:

$$P\left(\frac{\left|\frac{X}{N_w}-p\right|}{p}<\frac{\varepsilon}{p}\right)\geqslant\delta \tag{21}$$

由式(21)可知,使用频率越大的词,其经过训练以后的统计频率误差会越小。例如,如果在上式中,设可信度 $\delta=0.98$,$\varepsilon=4.375\times10^{-7}$,则对汉语中使用频率最高的"的"字,尽管在统计前无法确切地知道它的统计频率,但从已有的资料和统计中,能够粗略地估计出它的使用频率 $p>2.5\times10^{-2}$,将该统计频率和 δ、ε 的值代入式(21),就有 98%的把握保证"的"字统计频率的误差不会超过 $\varepsilon/p<4.375\times10^{-7}/(2.5\times10^{-2})=1.75\times10^{-5}$,而对那些使用频率较低的词,估计的相对误差就会大些。相对误差越小,利用 MLE 法所建立的语言模型的描述准确性就会越高。

结束语 由于 Unigram 模型是 n-gram 统计模型中的最简单一种,因此,它的训练语料样本规模可以看作是 n-gram 模型训练样本规模的下界,式(20)就是该下界的估计公式。它是考虑了建模语言单位在语言中的使用频率不同,其使用频率估计的误差要求就应该不同,得到的结果表明其更适于在实际中应用。

参 考 文 献

[1] 关毅. 基于统计的汉语语言模型研究[D]. 哈尔滨工业大学. 北京:国家图书馆,1999

[2] 刘爱芹. 随机抽样中样本容量确定的影响因素分析[J]. 山东财政学院学报,2006:60-64

[3] 张湘平,张金槐,谢红卫. 关于样本容量、验前信息与 Bayes 决策风险的若干讨论[J]. 电子学报,2003,31(4):536-538

[4] 王学民. 多指标分层抽样中样本容量折衷分配的加权方法[J]. 统计与决策,2008,3:27-29

[5] 耿修林. 整群抽样审计时样本容量的确定[J]. 审计研究,2007(6):85-88

[6] 盛骤,谢式千,潘承毅. 概率论与数理统计[M]. 北京:高等教育出版社,2001

[7] 北京语言学院语言研究所编. 现代汉语频率词典[M]. 北京:北京语言学院出版社,1986

[8] 刘小勤. 现代汉语分词词表的选词方法研究[D]. 太原:山西大学,1999