一种改进的面向文本的领域概念筛选算法

傅 鹂[1] 黄利强[2] 付春雷[3]

(重庆大学软件学院 重庆 400044)[1] (重庆大学数学与统计学院 重庆 401331)[2]

(重庆大学信息与网络管理中心 重庆 400044)[3]

摘 要 在语义技术及其应用中,本体学习是一个研究热点,而领域概念筛选则是本体学习的基础。对于领域概念筛选问题,领域一致度与领域相关度相结合的方法效果较好,却也存在信息描述不全的缺点,因此提出了一种针对此问题的改进的领域概念筛选算法。通过计算候选概念之间的语义相似度,识别出低频的具有同义关系和整体-部分关系的词语集,过滤掉部分冗余概念,然后采用改进的领域相关度和领域一致度相结合的公式进行筛选。实验表明,该方法提高了领域概念筛选的有效性。

关键词 语义技术,本体学习,领域概念,筛选算法,上下文

中法图分类号 TP18 **文献标识码** A

Improved Text-oriented Algorithm for the Domain-specific Concept Sieving

FU Li[1] HUANG Li-qiang[2] FU Chun-lei[3]

(School of Software Engineering, Chongqing University, Chongqing 400044, China)[1]

(College of Mathematics and Statistics, Chongqing University, Chongqing 401331, China)[2]

(Information and Network Management Center, Chongqing University, Chongqing 400044, China)[3]

Abstract Ontology learning is a hot research field in semantic technology and its application, and the domain-specific concept sieving is the foundation of ontology learning. Although the method based on domain relevance expressions and domain consensus expressions displays the good effectiveness for the domain-specific concept sieving, it exists the faults of unilateral description information. So this paper presented an improved sieving algorithm of the domain-specific concept to solve the above problems. First, low frequency with synonymy and the part-of relationship words set were identified and redundant concepts were filtered out through calculating the semantic similarity between candidate concept, and then using the improved field concept similarity and domain concepts consistent degree formula sieve concepts. The experiments show that this method improves the effectiveness of the domain-specific concept sieving.

Keywords Semantic technology, Ontology learning, Domain-specific concept, Sieving algorithm, Context

1 引言

万维网已是人们获取知识的主要来源,但从 Web 上准确搜索到所需信息并非易事,主要原因是当今的搜索引擎大都采用关键字匹配,未考虑语义信息,搜索结果包含大量不相关页面,同时遗漏大量相关页面。针对这个问题,万维网创始人 Tim Ber ners-Lee 提出了语义万维网(semantic Web)的构想[1]。语义万维网采用基于语义的多层次表示框架,其中本体(ontology)位于由数据描述到知识推理的转折层,因此本体的创建是实现语义网的关键环节。到目前为止,本体创建主要依赖人工方式,需要用户逐个输入和编辑每个概念的名字、属性、约束等内容,不仅耗时费力、容易出错,而且难以做到及时动态更新。因此,为实现本体自动或半自动构建的本体学习(ontology learning)技术应运而生。

Alexander Maedche 给出了本体的形式化定义[2],它是一个五元组 $O:=\{C,R,H^c,Rel,A^o\}$。其中,C 和 R 是两个不相交的集合,C 中的元素称为概念,R 中的元素称为关系;H^c 表示概念层次,即概念间的分类关系;Rel 表示概念间的非分类关系;A^o 表示本体公理。由此可见,本体学习的任务包括概念的获取、概念间关系(包括分类关系和非分类关系)的获取和公理的获取 3 个层面,其构成了从简单到复杂的层次。概念是本体最基本的组成元素,决定了下一阶段关系和公理获取的效果。所以,概念获取的准确率和召回率就显得尤为关键。为此,本文限制在本体的概念获取。另一方面,文本由于是万维网基本的、最主要的信息形式,因此对面向文本的概念获取方法的研究至关重要。

2 背景知识及相关工作

文本作为万维网最基本的、最主要的信息形式,成为当前本体学习最主要的数据源。而在面向文本的本体学习中,概

本文受重庆市科委自然科学基金计划重点项目(CSTC,2011BA2022)资助。

傅 鹂(1961-),男,教授,主要研究方向为信息安全、语义技术及其应用、知识库及应用等,E-mail:fuli@cqu.edu.cn;黄利强(1986-),男,硕士,主要研究方向为本体学习、语义技术;付春雷 男,工程师,主要研究方向为网络安全、本体学习、神经网络。

念获取是首要的一步，得到了广泛的研究，产生了很多方法。当前应用最广的方法主要有：基于语言学的方法[3]、基于统计的方法[4]、基于种子概念的方法[5]、混合方法[6]。这些方法多是首先获取领域中的候选概念(candidate concepts)，即领域知识中的一些词语；然后进行领域概念筛选，得到领域概念，即具有明显领域代表性且意义单一的概念。领域概念筛选的现有方法主要包括基于互信息、基于频率、基于领域相关度和领域一致度等几种，其中基于领域相关度与领域一致度相结合[7]的方法效果较好。

定义 1 领域相关度反映概念与领域的相关程度。设总共 n 个领域$\{D_1, D_2, \cdots, D_n\}$，则候选概念 c 对于某领域 D_i 的领域相关度为：

$$DR(c, D_i) = \frac{P(c|D_i)}{\max\limits_{1 \leqslant i \leqslant n} P(c|D_i)} \tag{1}$$

式中，条件概率 $P(c|D_i)$可用频率估计：

$$P(c|D_i) \approx \frac{f_{c,i}}{\sum\limits_{c' \in D_i} f_{c',i}} \tag{2}$$

式中，$f_{c,i}$表示候选概念 c 在领域 D_i 中的频数。

定义 2 领域一致度反映概念在领域中的分布均匀度。设领域 D_i 有 m 个文本文档$\{d_1, d_2, \cdots, d_m\}$，则候选概念 c 对于 D_i 的领域一致度为：

$$\begin{aligned} DC(c, D_i) &= H(c, D_i) \\ &= -\sum_{d_j \in D_i} P(c, d_j) \cdot \log_2 P(c, d_j) \end{aligned} \tag{3}$$

式中，联合概率 $P(c, d_j)$可用频率估计：

$$P(c, d_j) \approx \frac{f_{c,d_j}}{\sum\limits_{d_j \in D_i} f_{c,d_j}}$$

式中，f_{c,d_j} 表示候选概念 c 在领域文档 d_j 中出现的频数。在式(3)中，$H(c, D_i)$为信息熵，其值越大，表明 c 在各文档中的分布越均匀，也即 c 在各文档中出现的频数越一致。当 c 出现在各文档中的频数全相等时，$H(c, D_i)$达到最大值。

当候选概念与某领域的相关度和一致度都较大时，该候选概念就可能是该领域中的概念。因此，通过对二者的加权平均，可定义候选概念 c 对领域 D_i 的一个综合判定指标：

$$DW(c, D_i) = \alpha \cdot DR(c, D_i) + \beta \cdot DC(c, D_i) \tag{4}$$

并将 $DW(c, D_i)$大于等于某个阈值 θ，作为候选概念 c 筛选为领域概念的判定准则。阈值 θ 可根据概念提取的准确率和召回率要求设定，而权重 α 和 β 反映领域相关度和领域一致度在领域概念筛选中各自的贡献。文献[7]通过实验发现，α 取 0.9 左右、β 值取 0.25～0.35 较合适。

然而，采用式(4)进行领域概念筛选，一方面容易漏选某些具有同义关系或整体-部分关系的低频候选概念；另一方面又易将某些与领域无关的高频冗余概念错选为领域概念，导致概念提取的准确率和召回率同时受到不利影响[8,9]。鉴于此，文献[8]将候选概念上下文信息引入到领域概念筛选中，能滤掉部分冗余概念；文献[9]考虑了具有同义和整体-部分关系的低频词在筛选时易被遗漏的情况。但因二者均仅对某一方面的问题进行了改进，如未给出同义关系和整体-部分关系的量化计算，故结果仍有明显不足，包括仍有冗余概念较多的问题。

3 算法改进

在领域概念筛选中，以领域专家人工提取的结果作为领域概念标准，将采用传统方法自动筛选出来的结果与之相比较，则必然是正确的结果与标准的语义相似度较大，而冗余概念与标准的语义相似度较小。基于这个思想，为克服文献[8,9]的不足，本文首先利用候选概念上下文中的信息计算候选概念之间的语义相似度，根据计算结果识别出低频的具有同义关系和整体-部分关系的词语集，并过滤掉部分冗余概念；然后，采用文献[9]提出的改进的领域概念相似度和领域概念一致度公式进行筛选，以解决低频领域概念易遗漏的问题。

3.1 候选概念上下文

现实中，任何一个概念的正确语义都不能通过字面上以及简单组合它的各个部分而得到，必须考虑到那些额外的、隐含在上下文(语境)中的信息，借助语境信息来确定概念的语义。

候选概念的上下文，一般是在候选概念前后的一定范围[8]，这个固定的范围被称为窗口，表示为$[a, b]$，即候选概念左 a 个位置和右 b 个位置的范围。窗口中的词语被称为该候选概念的上下文词语，通常采用如下步骤选取：

1) 对各个领域文本文档以句为单位进行分词和词性标注；

2) 在句子中，以候选概念为中心，剔除句中无意义的虚词，选用实词(主要为名词、动词、形容词、副词)作为上下文词语，并计算各上下文词语的词频，形成关于该候选概念的上下文向量：

$$\boldsymbol{S}_i = [(C_1, W_1), (C_2, W_2), \cdots, (C_n, W_n)]$$

式中，$\boldsymbol{S}_i$ 为候选概念的上下文向量，C_i 为上下文词语，W_i 为 C_i 在候选概念窗口内出现的次数；

3) 对领域中的任意两个候选概念，分别提取上下文向量，采用余弦法计算语义相似度：

$$\mathrm{sim}(\boldsymbol{S}_i, \boldsymbol{S}_j) = \cos(\boldsymbol{S}_i, \boldsymbol{S}_j) = \frac{\sum\limits_{i=1}^{n} W_i X_i}{\sqrt{\sum\limits_{i=1}^{n} W_i^2} \sqrt{\sum\limits_{i=1}^{n} X_i^2}} \tag{5}$$

式中，W_i 表示一个候选概念上下文向量中第 i 个词语的词频，X_i 表示另一个候选概念上下文向量中第 i 个词语的词频。

3.2 筛选准则的改进

在进行领域概念筛选时，很多重要的低频领域概念易被漏掉。比如，与某领域概念具有同义关系或整体-部分关系的那些领域概念，可能由于在领域中出现的频数较低，因而领域相关度和领域一致度较小，结果被错误地排除掉，降低了概念获取的准确率和召回率。因此，如何提取出这些低频的领域概念就显得尤为重要。下面分别从同义关系和整体-部分关系分析入手，寻求到更好的筛选准则。

一个领域概念可以由多个具有同义关系的词语代表，或等价地，由一个同义词集代表。如果同义词集较大(即同义词数量较多)，就很可能造成每个词语在领域文档中出现的频数相对较小，从而导致该领域概念易被漏选。另一方面，所有具有同义关系的词语在文档中的语境信息应是类似的，所以它们两两之间的语义相似度应较大且彼此较为接近。根据这一特征，可以找出同义词集，用以改进筛选准则，使得在领域文档中出现频率较低的概念不易被漏掉，具体做法是：

设有一词集 T 为候选概念集的子集。若 T 中两两词语的相似度大于等于某个 θ_1，同时每个词语与不属于 T 的词语

的相似度小于θ_1，则 T 将作为可能代表某个领域概念的同义词集。

整体-部分关系，比上述同义关系更复杂。在分析了大量的领域文档之后，发现领域中出现的绝大多数整体词与部分词具有以下主要特征：

1) 整体词与各个部分词的语义相似度彼此较接近；

2) 整体词与非部分词的语义相似度较小；

3) 整体词出现的频数较小，语境信息缺乏；

4) 各部分词彼此不重叠，语境信息差别较大。

可以看到，与同义关系的情况类似，这里也遇到与领域相关的低频词语，很可能被常规筛选方法排除在外。为改进筛选策略，以期获得更好的结果，首先要找出具有整体-部分关系的词集。综合整体词与部分词的特征，具体做法如下：

设有一词集 M 为候选概念集的子集。考察不属于 M 的某个候选概念 c。若 c 与 M 中每个词语的相似度都大于等于某个阈值 θ_2，与不属于 M 的词语的相似度小于 θ_2，同时，M 中两两词语的相似度值波动较大，则 M 将作为对应于 c 的整体-部分关系词集。

上述用到的 θ_1 和 θ_2 可根据提取概念要求的准确率和召回率具体确定。

考虑了同义关系和整体-部分关系情况下的低频词因素之后，领域相关度和领域一致度的计算可做如下相应改进。

对式(1)中分子 $P(c|D_i)$ 的改进分以下几种情况：

1) 若候选概念 c 有同义词，且 c 与其它候选概念不存在整体-部分关系时，设与 c 有同义关系的候选概念集合为 $T=\{s_1,s_2,\cdots,s_n\}$，则有：

$$P(c|D_i)=\sum_{s_j\in T}P(s_j|D_i) \tag{6}$$

2) 若候选概念 c 与候选概念集 $M=\{p_1,p_2,\cdots,p_n\}$ 存在整体-部分关系，且不存在同义词时，则有：

$$P(c|D_i)=\sum_{p_j\in M}P(p_j|D_i) \tag{7}$$

3) 若候选概念 c 与候选概念集 $M=\{p_1,p_2,\cdots,p_n\}$ 存在整体-部分关系，且与候选概念集 $T=\{s_1,s_2,\cdots,s_n\}$ 存在同义词关系，则有：

$$P(c|D_i)=\sum_{p_j\in M}P(p_j|D_i)+\sum_{s_j\in T}P(s_j|D_i) \tag{8}$$

同理，可采用同样的方法对式(3)中的 $P(c,d_j)$ 进行改进。

利用改进公式计算得到候选概念的领域相关度和一致度，将其带入式(4)判断 $DW(c,D_i)$ 是否大于等于预先设定的阈值，以筛选领域概念。

4 算法描述

下面给出改进的概念筛选算法描述。

输入：候选概念集合

输出：领域概念集合

1) 设有 n 个领域的领域集合 $set_D=\{D_1,D_2,\cdots,D_n\}$，其中 D_i 为领域语料库，其它的为过滤语料库，候选概念集 $set_{cc}=\{c_1,c_2,\cdots,c_n,\cdots c_t\}$，对 set_{cc} 中的每个候选概念运用式(5)进行相似度计算，根据计算结果值删除部分冗余概念，然后结合同义词和整体-部分词的判别准则，对候选概念 c_i 执行 2)－8)，直到所有的候选概念处理完毕。

2) 如果候选概念 c_i 与其它候选概念既不存在同义词关系也不存在整体-部分关系，则直接利用式(4)计算。

3) 如果候选概念 c_i 在 set_{cc} 中只存在同义词，则找出它的同义词集，利用式(6)计算 $P(c_i|D_i)$。

4) 如果候选概念 c_i 在 set_{cc} 中只存在整体-部分关系词，则找出它的部分词集，利用式(7)计算 $P(c_i|D_i)$。

5) 如果候选概念 c_i 在 set_{cc} 中既有同义词又有整体-部分关系词，则找出它的同义词集和部分词集，利用式(8)计算 $P(c_i|D_i)$。

6) 若候选概念 c_i 在 set_{cc} 中与其它候选概念满足 3)－5)中的一种，则将相应的 $P(c_i|D_i)$ 的计算结果代入到式(1)，得到候选概念 c_i 的领域相关度。

7) 同理，计算 $P(c,d_j)$ 时与 3)－5)用到的方法类似，将计算结果带入式(3)，得到候选概念 c_i 的领域一致度。

8) 将 6)和 7)的结果代入到式(4)进行计算，得到候选概念 c_i 的判断结果 $DW(c_i,D_i)$。

9) 根据 8)的结果进行判断，如果候选概念 c_i 的判断结果 $DW(c_i,D_i)$ 大于阈值 θ，则认为概念 c_i 是领域 D_i 上的领域概念；否则，认为 c_i 不是领域 D_i 上的领域概念，从候选概念集合 set_{cc} 中删除概念 c_i。

10) 将领域概念集合 set_{cc} 作为结果返回。

5 数值试验及比较

为了证明所提方法的有效性，本文与文献[8,9]提供的领域概念筛选方法在同一数据集上进行了对比实验。实验数据集采用复旦大学文本语料库的一部分，包括计算机 134 篇、环境 134 篇、交通 143 篇、教育 147 篇、军事 166 篇、艺术 166 篇、政治 338 篇，总共 7 个类别、1228 篇文本文档。将军事作为领域语料，其他几类作为过滤语料。按照上述领域概念筛选算法，使用准确率(precision)、召回率(recall)、综合测量值($F_{measure}$)3 个指标[10]来与文献[8]进行对比实验。3 个指标的计算方法如下：

$$precision=\frac{correct_{extracted}}{all_{extracted}}$$

$$recall=\frac{correct_{extracted}}{all_{corpus}}$$

$$F_{measure}=\frac{2precision\cdot recall}{precision+recall}$$

式中，$correct_{extracted}$ 表示筛选得到的正确的领域概念数，$all_{extracted}$ 表示筛选得到的所有的概念数，all_{corpus} 语料库中所有的概念数。

通过与文献[8]提出的最新概念筛选算法比较得到实验结果，如表 1 所列。

表 1 数值实验比较

参数和指标	改进前	改进后
文档数/过滤文档数	20/10	20/10
all_{corpus}	184	184
$all_{extracted}$	81	106
$correct_{extracted}$	62	88
precision	33.7%	47.8%
recall	76.5%	83%
$F_{measure}$	46.79%	60.67%

从表 1 可以看出，改进后的方法对领域概念提取的准确率和召回率都有很大程度的提高，说明改进后的算法对领域概念中的低频同义词、整体-部分关系词、同义且具有整体-部分关系词的筛选效果十分显著；同时，在这些低频词中有很多都属于领域概念，避免了一些非常重要的低频领域概念被筛选掉。

同样，改进后的算法较文献[9]的算法有以下几点优势：

1）改进后的算法使用相似度删除了部分冗余概念，对一般概念提取的准确率较高。

2）根据候选概念上下文的语义情境给出了低频同义词和整体-部分关系词的判别准则，在实际应用中执行力更强。

结束语 本文同时考虑了因为同义关系和整体-部分关系低频词带来的领域概念筛选问题，通过更细致的考虑改进了原来的计算公式，并与现有方法进行了实验比较，表明改进后的方法有明显效果。

参 考 文 献

[1] 朱礼军，陶兰，黄赤. 语义万维网的概念、方法及应用[J]. 计算机工程与应用，2004，3：79-84

[2] Maedche A. Staab S. Ontology learning for the semantic Web [J]. IEEE Intelligent systems，2001，16(2)：72-79

[3] Shamsfard M，Barforoush A. Learning ontologies from natural language texts[J]. International Journal Human-computer Studies，2004，60(1)：17-63

[4] Missikoff M，Navigli R，Velardi P. Integrated approach for Web ontology learning and enginee-ring[J]. IEEE Computer，2002，35(11)：60-63

[5] Moldovan D，Girju R，Rus V. Domain-specific knowledge acquisition from text [C]//Proc. of The Sixth Conference on Applied Natural Language Processing. 2000：268-275

[6] Velardi P，Fabriani P，Missikoff M. Using text processing techniques to automatically enrich a Domain ontology [C]//Proc. of the Formal Ontology in Information Systems. 2001：270-284

[7] 贾秀玲，文敦伟. 面向文本的本体学习中概念提取及关系提取的研究[J]. 中南大学，2007

[8] 张玉芳，杨芬，熊忠阳. 基于上下文的领域本体概念和关系的提取[J]. 计算机应用研究，2010

[9] 王红滨，刘大昕，王念滨. 一种本体学习中的领域概念筛选算法[J]. 系统工程与电子技术，2010

[10] Liu Bai-song，Gao Ji. General ontology learning frame-work[J]. Journal of Southeast University(English Edition)，2006，22(3)：381-384

（上接第 222 页）

运行中时间较长，由于调度器处于学习的阶段，因此运行的效率比较低。之后作业的运行时间就处于一个稳定的状态，运行时间收敛。

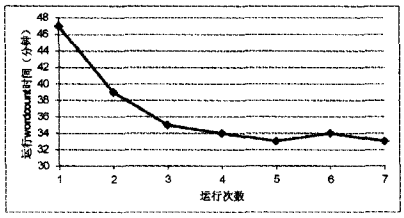

图 3 运行 WordCount 作业与运行次数关系图

为了观察特征加权对 WordCount 作业运行时间的影响，将 CPU、内存、磁盘、网络的 $SGF(f_i, D)$ 值设为相同，如都设为 4。此时根据公式计算，CPU、内存、磁盘、网络的权值都为 1，简化为不带权值的朴素贝叶斯分类调度。同样重复运行 7 次，WordCount 作业的两种比较如图 4 所示。

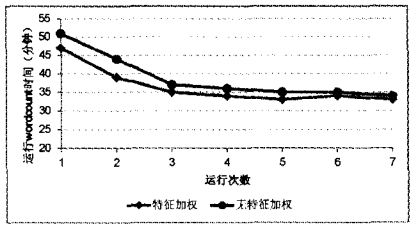

图 4 特征加权与无特征加权作业比较

从图 4 中可以看出，在对资源估计值都相同的条件下，通过调整特征变量的权值可以缩短作业运行的时间，有效地提高资源的使用效率。特别是在作业运行时，头 2 次作业运行时间缩短比较明显；通过学习之后，特征加权和无特征加权趋近相同，但总体上特征加权之后效率得到提高。

参 考 文 献

[1] Cloud computing [EB/OL]. http://en. wikipedia. org/wiki/Cloud_computing，2010-05-10

[2] 刘鹏. 云计算[M]. 北京：电子工业出版社，2010

[3] MapReduce[EB/OL]. http://zh. wikipedia. org/zh-cn/MapReduce，2010-03-05

[4] Ghemawat S，Gobioff H，Leung S T. The Google file system[J]. ACM SIGOPS Operating Systems Review，2003，37(5)：29-43

[5] 中科院计算所网络重点实验室. Hadoop 作业调度研究[EB/OL]. http://www. slideshare. net/YongqiangHe/hadoopv01，2010

[6] Zaharia M. Job Scheduling with the Fair and Capacity Schedulers [EB/OL]. http://www. cs. berkeley. edu/～matei/talks/2009/hadoop_summit_fair_scheduler. pdf，2009-07-10

[7] Capacity Scheduler Guide[EB/OL]. http://hadoop. apache. org/common/docs/r0. 20. 2/capacity_scheduler. html，2010

[8] 王峰. Hadoop 集群作业的调度算法[J]. 程序员，2009，12

[9] Fair Scheduler Guide[EB/OL]. http://hadoop. apache. org/common/docs/r0. 20. 2/fair_scheduler. html，2010-01-02

[10] Zaharia M，Konwinski A，Joseph A D，et al. Improving mapreduce performance in heterogeneous environments[C]//USENIX Association. 2008：29-42

[11] Polo J，Carrera D，Becerra Y，et al. Performance-driven task co-scheduling for mapreduce environments[C]//12th IEEE/IFIP Network Opertions and Management Symposium. 2010：373-380

[12] Kc K，Anyanwu K. Scheduling Hadoop Jobs to Meet Deadlines [C]//CLOUDCOM'10 Proceedings of the 2010 IEEE Second International Conference on Cloud Computing Technology and Science. 2010：388-392

[13] Sandholm T，Lai K. Dynamic proportional share scheduling in Hadoop[C]//Springer. 2010：110-131

[14] Kunz T. The Learning Behaviour of a Scheduler using a Stochastic Learning Automation[J]. Relation，1991，10(1-127)：2600

[15] Negi A，Kishore K. Applying machine learning techniques to improve linux process scheduling[C]//IEEE. 2005：1-6

[16] Santos L P，Proenca A. Scheduling under conditions of uncertainty：a bayesian approach[C]//Springer. 2004：222-229

[17] Zhang H. The optimality of naive Bayes[J]. A A，2004，1(2)：3

[18] 邓维斌，王国胤，王燕. 基于 Rough Set 的加权朴素贝叶斯分类算法[J]. 计算机科学，2007，34(002)：204-206