基于词频分布变化统计的术语抽取方法

周 浪[1,2] 张 亮[3] 冯 冲[2] 黄河燕[2]
(南京理工大学计算机科学与技术学院 南京 210094)[1] (计算机语言信息工程研究中心 北京 100089)[2]
(南京大学计算机科学与技术学院 南京 210093)[3]

摘 要 提出了一种规则与统计相结合的术语抽取方法,用于抽取包含多个词语的词组型术语。目前,绝大多数的统计方法都侧重于衡量术语的结构完整性,但这些方法并不能体现术语与专业相关的领域特征。通过对术语在各文档中的分布情况进行观察,提出了一种利用术语在语料中词频分布变化程度的统计信息来检验术语的领域相关性的方法,同时结合机器学习方法获取的语言知识,从计算机领域的语料中抽取领域特征明显的词组型术语。实验证明,该方法对低频术语和高频普通词串有较强的分辨能力。

关键词 术语抽取,机器学习,分布方差,知识获取,termhood,unithood

Terminology Extraction Based on Statistical Word Frequency Distribution Variety

ZHOU Lang[1,2] ZHANG Liang[2] FENG Chong[2] HUANG He-yan[2]
(College of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China)[1]
(Research Center of Computer & Language Information Engineering, CAS Beijing 100089, China)[2]
(Dept. of Computer Science and Technology, Nanjing University, Nanjing 210093, China)[3]

Abstract A hybrid terminology extraction system combined with linguistic knowledge and statistical information was introduced to extract compound terms which contain more than one word. There have been many statistical strategies used in automatic terminology extraction, most of which emphasize particularly to measure the integrality of the terms, other than domain features. To measure the domain relativity of terms, a mew method utilizing term frequency distribution variety was proposed. Incorporating with linguistic knowledge acquired by machine learning method, an automatic extraction system was implemented to extract multi-word terms from the corporate of computer domain. The results show that this approach is effective especially to distinguish terms with lower frequency and common words with higher frequency.

Keywords Terminology extraction, Machine learning, Distribution variance, Knowledge acquisition, Termhood, Unithood

1 引言

近几十年来,在经济、文化、科技高速发展的过程中产生了越来越多的新技术、新产品及新的概念,随之变化最大的就是各学科领域内的术语。术语是这些学科知识的集中体现,如果要了解这些学科的发展动态,接触其中的术语是不可避免的。而旧式的依靠人工收集这些术语显然跟不上更新的速度,而且需要耗费大量的人力物力。与此同时,计算机技术不论是在软硬件上都得到了快速的提升,因此,利用计算机技术来实现术语自动抽取这一需求也就应运而生。术语自动抽取技术的应用范围非常广泛,不仅仅可以用来编撰专业辞典,还可以应用于信息检索、命名实体的识别、机器翻译、自动生成文摘等领域。

现有的术语自动抽取方法主要分为三种:(1)基于语言规则的方法[1];(2)基于统计信息的方法[2];(3)规则与统计相结合的方法[3]。由于规则的覆盖面小,构造规则库十分耗费人力物力;而纯统计的方法中统计模型计算的准确性主要依赖于语料库的规模,语料库增大的同时引发的数据稀疏问题很难解决。因此,规则与统计相结合的方法是目前的研究趋势。Kageura 在文献[4]中将衡量术语的统计标准又细分为两种:一是表示术语作为一个独立的语言单位,其语言结构应该是稳固的,称为 unithood;二是术语作为一个领域知识的代表,负载着很大的信息量,应与领域知识密切相关,称为 termhood。目前,绝大多数的统计量都是针对术语的第一个特性 unithood 的,如:互信息、log-likelihood、左右熵等,都是统计术语内部词语的结合度,按 unithood 值的大小进行排序,或设定阈值进行过滤。从严格意义上来讲,这些方法抽取出来的只能称为短语,几乎未能体现术语的 termhood 特性。

到稿日期:2008-06-24 本文受国家 863 高技术研究发展计划项目(2006AA01Z152),国家自然科学基金项目(60672149)资助。
周 浪(1982—),女,博士生,CCF 会员,主要研究领域是自然语言处理、自动术语识别,E-mail:yzzhoulang@126.com;张 亮(1966—),男,博士,主要研究领域是自然语言处理、自动问答系统;冯 冲(1977—),男,博士,主要研究领域是自然语言处理、命名实体识别;黄河燕(1963—),女,博士生导师,主要研究领域是自然语言处理、机器翻译。

也有少部分研究是针对 termhood 展开的。文献[5]中使用了基于 TFIDF 方法来抽取专业词汇，除了专业领域的语料外(前景语料)，还使用了另一种专业的语料(背景语料)，统计两个语料中术语的词频对比变化作为衡量术语 termhood 的标准。文中使用的计算方法过分依赖于使用的背景语料，如果两个语料库的专业交叉性不大，则对其前景语料中高频的普通词汇识别能力不够，如："实验数据"在科技文章中可能普遍存在，但是在体育领域出现的概率要小很多；反之，如果两个语料的专业有一定交叉，又会削弱术语的识别能力；文献[6]中除了使用互信息和 log-likelihood 来衡量术语内部词语之间的结合能力外，还利用了 CBC 聚类方法从抽取的术语文本中自动剔除非术语的候选项。但是该方法需要人为设定一些术语作为种子术语，而且容易将在某一篇文档中频繁出现的术语和种子术语进行合并。

通过观察术语在语料中的词频分布特性，发现术语在每篇文档中出现的频次变化较大。籍此，本文提出一种通过衡量术语在语料中词频分布变化程度的 termhood 计算方法。该方法相对以上几种方法的改进主要有：

(1)对低频术语的识别能力较强；

(2)能够排除掉大部分语料中的普通词串，包括那些出现频率很高的词串；

(3)所需的语料资源较少，而且在语料规模较小的情况下，也能够达到令人满意的正确率。

本文除了统计术语的词频分布变化程度外，同时还结合了利用机器学习方法从术语资源中获取的一系列语言知识，构成一个完整的中文术语抽取系统，主要用于抽取语料中的词组型术语。在下面的章节中，将会详细介绍该系统的工作流程及原理。

2 基于分布变化统计的术语抽取系统设计

本文实现了一个语言规则和统计信息相结合的中文术语抽取系统，该系统主要由 3 个模块组成，其结构流程如图 1 所示。

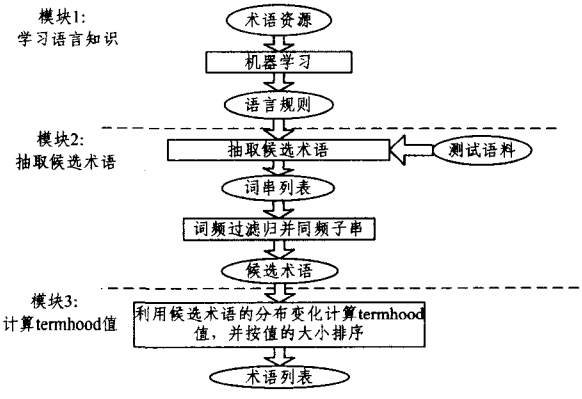

图 1 基于术语分布变化的术语抽取系统

学习语言知识：使用中科院计算所词法分析系统 ICTCLAS 对人工收集的术语进行分词及词性标注，并利用机器学习的方法从中获取一系列的语言规则。

抽取候选术语：同样使用 ICTCLAS 对测试语料进行词性标注，依据上一个模块中获取的语言规则及一个停用词表抽取出一个词串列表。与此同时，统计各个词串的词频、文档频率及在每篇文档中出现的频次。只选取词频大于 2、归并同频子串后的词串进入候选术语列表。

计算 termhood 值：利用候选术语在语料中词频的分布变化作为指标，计算其 termhood 值。并按 termhood 值的大小，由高到低进行排序，构成最终的术语列表。

在下面的几个章节将详细介绍模块 1 和模块 3 的执行过程及细节。最后通过实验，将本文提出的方法和 C-value 方法及互信息方法进行对比，并分析实验结果。

3 语言知识的获取

术语作为自然语言的一种表现形式，能够独立表达专业知识，因此其语法结构必然是稳固的。在前面介绍的方法中，很多研究都是抽取名词短语作为候选术语。但事实上，很多术语属于动词短语，如"归并排序"、"调页"等等。而且，现有的词性标注系统的正确率并不能达到 100%，其中还存在不少的错误，此时再以人工编撰的规则来抽取术语，只会削弱术语识别的正确率和召回率。因此，本文在抽取候选术语时，不会将抽取的规则限制在名词短语或人工编撰的规则这一狭小的范围内。

在获取语言知识之前，需要对已有的术语资源进行预处理。本文使用中科院计算所的自动分词系统 ICTCLAS 对已有的术语资源进行分词及词性标注。之后，并利用机器学习的方法从中获取一系列的术语语法、结构和组成特征，在下个阶段将这些特征应用在候选术语的抽取过程中。

3.1 术语资源

在科技文档中，术语的出现频次很高，尤其是文章中由作者设定的关键词，都是最能体现文章主要内容、主题的术语。我们收集了 04 年－07 年间《计算机学报》、《计算机研究与发展》、《软件学报》和《中文信息学报》中所有文章的关键词，去掉重复项后，共有 10,026 条关键词。在本文的工作中，将这些关键词视为最能体现领域特征的标准术语，对其结构成分进行分析、学习，总结出术语的结构、组成特征模式。

3.2 术语的长度特征

本文将术语的长度定义为术语中包含的词语数，即分词后分成的词语数。对这一特征进行分析后发现，术语的长度变化范围在 1～10 之间。其中，长度在 2～6 之间的术语数目最多，有 8,871 条，占总数的 88.48%。长度为 1 的术语有 1,089 条，占10.86%，这其中有 445 条都是英文缩写。而长度在 7～10 之间的术语只有 66 条，占总数的 0.66%。

张普在报告[7]中指出，中文术语 2～6 个字的占大多数，为 76.9%。张榕[8]在对包含 328,150 条术语的术语数据库分析后，也发现术语的长度一般以 2,3,4 居多，占总数的 71.723%，大部分的术语长度在 1～6 之间，大于 6 的仅有 0.572%，文献[8]中的统计结果和本文的完全相符。因此在下面候选术语工作中，只抽取长度在 2～6 之间的词串作为术语候选词。

3.3 术语的语法结构特征

对长度在 2～6 的术语进行词性标注后，共切分为 24,127 个词，发现非语素词、语气词和状态词没有出现过，而叹词、成语、拟声词、代词、处所词和标点符号只出现了 67 次。在术语的第一个词语中，助词、连词、后接成分也出现得很少，同样，末尾的词中，前接成分、方位词、连词和助词很少出现。另外，还发现包含名词、动词、量词、后接成分、习用语或简称

略语的术语占了 99.74%。

根据以上的观察和统计，本文制定四条术语候选词的抽取规则：规则一，术语中不包含叹词、成语、代词、处所词、标点符号、非语素词、语气词和状态词；规则二，术语不得以词性为助词、连词或后接成分的词开头；规则三，术语不得以词性为前街成分、方位词、连词或助词的词结尾；规则四，术语中至少包含下列词性中的一种：名词、动词、量词、后接成分、习用语、简称略语。

在下一个模块中，主要使用长度限制及这四个规则来抽取出候选术语，这相对于文献[2,3]中使用的语言知识要宽松很多。

4 基于术语词频分布变化统计的 termhood 计算方法

通过统计术语在语料中的词频分布信息可以发现，大多数术语，如“句法分析”、“机器学习”等，不论其总的词频数是高或是低，或是在语料中覆盖率如何，但术语在每篇文档中的出现频次差距较大；而普通的短语，如“实验数据”、“研究人员”等，在语料中的分布则比较稳定，即使其总的出现频次较高、出现的文档数不多，但在每篇文档中的出现频次变化却不大。图 1 反映了两个术语“句法分析”和“实验数据”在语料每篇文档中的词频变化对比。

为了更直观地进行比较，本文对术语的词频进行了归一化处理，采用出现的词频比例代替绝对的频次，如“句法分析”在整个语料中共出现了 309 次，而第一篇文档中出现了 11 次，占其在整个语料中出现次数的 3.56%。而且如果考虑到词频为 0 的点，那些出现的文档频次较少的普通词串的变化曲线也会不断呈现小幅度的抖动变化，所以去除了词频为 0 的点，使得普通词串在文档中出现频次的稳定性更加突出。

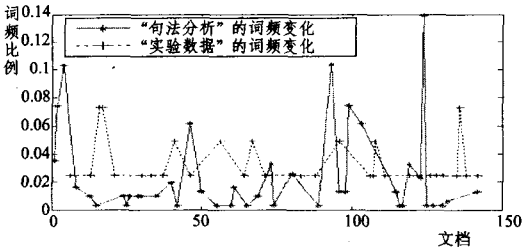

图 2 术语“句法分析”和“实验数据”在文档中的词频变化

通过图 2 的对比，不难发现术语在文档中的词频变化比较大，曲线抖动相对较为剧烈。而普通短语的出现则比较平稳，上下浮动不大。分析语料发现，在科技文献中，术语的出现一般分为两种情况：(1)文档的主要内容和该术语关系密切，则该术语被提及的次数很频繁；(2)文档与该术语属于同一类别内，但并不是直接相关，所以会有所提及，但次数较少。正因为如此，术语在不同的文档中，出现的词频才会有较大的变化。可见，词频分布的变化能对鉴别一个候选项是否为专业术语做出重要的指示。本文正是基于这种观察，提出了基于词频分布变化的 termhood 计算方法。

检验样本和总体分布的波动程度，最直接有效的方法就是利用样本方差。当方差的值越小，表示这个样本或总体的波动越小，也就是变化越平稳。假设有候选术语 t，基于样本分布变化的 termhood 计算方法如下所示：

$$DV\text{-}termhood(t)=\frac{tf(t)}{df(t)}\cdot\sigma=\frac{tf(t)}{df(t)}\cdot\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(tf_i(t)-\overline{tf}(t))^2}$$

其中，$tf(t)$表示候选术语 t 在整个测试语料中出现的总频率；$df(t)$表示候选术语 t 出现的文档频率；N 表示包含候选术语 t 的文档数；$tf_i(t)$表示候选术语 t 在第 i 篇文档中出现的频率；$\overline{tf}(t)$表示候选术语 t 在 N 篇文档中出现的平均频率。

此外，考虑到部分术语的文档频次非常小，如：“fisher 线性判别式”只在 1 篇文档中出现过 12 次，需要对上式进行修正。在每个术语的分布中引入一个均值点，使得术语的文档频次增加 1，记新的文档频次为 $N+1$，增加的这个点为$(N+1,tf^*)$，tf^* 表示候选术语 t 修正后在整个语料中出现的平均频率。则上式修正后如下所示：

$$DV\text{-}termhood(t)=\frac{tf(t)}{df(t)}\sqrt{\frac{1}{N}\sum_{i=1}^{N+1}(tf_i(t)-\overline{tf}^*(t))^2}$$

设语料中共包含 M 篇文档，则$\overline{tf}^*(t)$可以通过下式计算得到：

$$\overline{tf}^*(t)=\frac{tf(t)+\frac{tf(t)}{M}}{N+1}=\frac{(M+1)tf(t)}{M(N+1)}$$

通过上式，可以看出，当一个候选术语出现的次数越多、涉及的文档数越少、在每篇文档中出现的次数相差越大时，就越可能是术语，这与上文提到的观察现象相符。下面，本文将通过实验来检验该方法对专业术语和普通词串的识别区分能力。

5 实验及结果分析

5.1 测试语料

实验使用的测试语料由 142 篇科技论文组成，抽取自 2007 年《中文信息学报》和《计算机学报》，去除其中的图表及公式后，共 1.27M。使用中科院计算所的汉语词法分析系统 ICTCLAS 进行分词及词性标注后，共 3.04M。分词后，语料中共包含 422,792 个词语。

5.2 实验结果及评估

经过模块 1 和模块 2 的处理之后，共从测试语料中抽取出 10,413 条候选术语，在模块 3 中，将这些候选术语按其 DV-termhood 值高低进行排序。目前，对实验结果进行评估的标准主要有正确率和召回率：

$$\text{正确率}=\frac{\text{正确的术语数}}{\text{抽取出的术语总数}}\times 100\%$$

$$\text{召回率}=\frac{\text{正确的术语数}}{\text{语料中包含的术语数}}\times 100\%$$

评估实验结果的正确率时，选取术语列表中前 2000 个术语，人工判断其正确与否。同时随机从语料中抽取出 50 篇文档，人工识别出其中包含的术语，共 1265 条，来检验实验结果的召回率。为了更好地对实验结果进行分析，下面将本系统的结果，同 C-value 方法和互信息方法的结果进行对比。实验结果及对比如表 1 和表 2 所列。

表 1 DV 方法、C-value 方法和互信息方法实验结果中前 2000 条术语的正确率

	正确率					
	Top100	Top200	Top500	Top800	Top1000	Top2000
DV	96.0%	92.5%	90.4%	89.3%	89.1%	79.6%
C-value	69.0%	65.5%	64.0%	66.1%	65.1%	57.6%
MI	62.0%	48.0%	46.2%	44.8%	44.9%	38.2%

表 2 DV 方法、C-value 方法和互信息方法实验结果的召回率

	召回率		
	Top1000	Top2000	Top5000
DV	15.97%	24.19%	38.89%

C-value	16.13%	23.95%	37.47%
MI	6.09%	12.56%	31.62%

5.3 结果分析

表1将3种方法实验中前2000个结果的正确率进行了比较，很明显可以看出，DV的正确率比其他两个方法要高出很多。在抽出的前1000个术语中，DV方法的正确率比C-value方法要高出24个百分点，比MI方法高出34个百分点。对3个实验抽取出的前1000个术语中错误的结果进行分析后发现，结构不完整的词串占了比较大的部分，如"中随机抽取"、"数据集上"，这类错误的词串一般以介词"中"、"上"、"基于"或动词"使用"、"进行"等开头或结尾。出现这种错误主要是由于本文使用的语言规则比较宽松，并没有像文献[9]中限制于名词短语。虽然这种错误在3个实验结果中都出现了，但具体所占的比例却有很大的差距，如表3所列。

表3 DV、C-value和MI方法实验抽取出的前1000个术语中错误词串结构完整和不完整的数目

	DV	C-value	MI
结构不完整	69	275	100
结构完整	39	74	346

由表3可以发现，DV和C-value结果的错误词串中，结构不完整因素占了绝大多数，DV为63.89%，C-value方法为79.80%。而MI方法则完全相反，结构不完整的词串只占了22.42%。虽然DV方法的错误词串中结构不完整因素占了比较大的比例，但是总的数量却远比另外两种方法少，由此可见，DV方法具有一定的能力能够排除结构不完整的词串。

对结构完整的错误词串进行分析后发现，3种方法的错误原因各不相同。在DV的实验结果中，那些词频较高，但文档频率非常低的普通词串会被误识为术语，如"网络聊天"和"新闻报道"，前者只在1篇文档中出现过，但词频高达73；后者在4篇文档中共出现了32次，这主要是由于作者采用了网络聊天或新闻报道的内容作为语料，这和本文使用的测试语料规模较小有关。而C-value方法中，普通词串的识别能力很差，错误中甚至包括最常用的"表1"、"图2"等，排名分别为48和93，但在其他两种方法中，这两个词的排名均在3000之后。而且C-value方法对低频术语的识别能力较弱，如："贝叶斯算法"，在语料中只出现了9次，在C-value方法的结果中排名2100，但是在DV的实验结果中排名675。MI方法则对那些低频、结构稳定的普通词串鉴别能力较弱，如变化较少的人名、地名和机构名等。

通过表2可以发现3种方法的召回率普遍都比较低，相差并不大。导致召回率偏低的主要因素是测试语料的规模很小，很多术语在整个测试语料中只出现了1，2次，在抽取候选术语模块中，就已经被过滤掉。除此以外，还有词性标注错误、语法规则覆盖面不够广等原因。

结束语 本文提出了一种利用术语在语料中的词频分布变化来衡量其termhood的计算方法，并结合一系列的语言知识，构成一个完整的中文术语抽取系统。从实验可以看出，在小规模的语料中，该方法要优于目前常用的C-value方法和互信息方法，尤其是具有较强的低频术语和高频普通词串识别能力。

在下一步的工作中，我们将针对实验中出现的错误，加入词串结构稳定性的检验，进一步优化并完善现有的术语抽取系统。

参 考 文 献

[1] Bourigault D. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases[C]//Proceedings of COLING'92. 1992:977-981

[2] Pantel P, Lin D. A Statistical Corpora-based Term Extractor[C]//Lecture Notes in Artificial Intelligence. Springer, Verlag, 2001:34-46

[3] Frantzi K T, Ananiadou S, Mima H. Automatic Recognition of Multi-word terms: the C-value/NC-value Method[J]. International Journal on Digital Libraries, 2000, 3(2):115-130

[4] Kageura K, Umino B. Methods of Automatic Term Recognition: A Review[J]. Terminology, 1996, 3(2):259-289

[5] 刘桐菊，于浩，杨沐昀. 基于TFIDF的专业领域词汇获取的研究[C]//第一届学生计算语言学研讨会论文集. 2002

[6] 李勇. 基于聚类方法对特定领域术语的自动筛选[J]. 计算机工程与科学，2008，30(2)：64-66

[7] 张普. 信息领域汉语术语的特征及其在语料中的分布规律[J]. 语言教学与研究，2001

[8] 张榕. 术语定义抽取、聚类与术语识别研究[D]. 北京：北京语言大学，2006

(上接第123页)

分析，而随着构件技术的快速发展，利用构件设计大型复杂软件系统的软件开发方法日趋成熟，如何评估构件软件的可靠性，研究构件软件的老化原因，设计针对构件软件的抗衰策略，保持构件软件的性能，都需要新的适合构件软件的研究方法。

本文利用马尔科夫模型，通过分析构件软件的可靠性，提出了一种能够评估构件软件的可靠性并通过软件抗衰来保持软件性能的软件分析方法，今后我们将结合实例来进一步研究构件软件抗衰问题。

参 考 文 献

[1] Avritzer A, Weyuker E. Monitoring smoothly degrading systems for increased dependability[J]. Empirical software Engng, 1997(2):55-77

[2] Dohi T, Goseva-Popstojanova K, Trivedi K S. Estimating software rejuvenation schedules in high-assurance systems[J]. Compute, 2001, 44:473-82

[3] Wei X, Yiguang H, Trivedi K S. Analysis of a two-level software rejuvenation policy[J]. Reliability engineering & system safety, 2005, 87(1):13-22

[4] Cross D, Harris C N. Fundamentals of Queuing Theory. John Wiley and Sons, New York

[5] Goševa-Popstojanova K, Trivedi K S. Architecture-based approach to reliability assessment of software systems[J]. Performance Evaluation, 2001, 45(7):179-204

[6] Barlow R E, Proschan F. Statistical theory of reliability and life testing: probability models. New York: Holt, Rinehart and Winston, 1975

[7] Yacoub S, Cukic B, Ammar H. A Scenario-based Reliability Analysis Approach for Component-based Software [J]. IEEE Transactions on Reliability, 2004, 53(4):465-480

[8] 毛晓光，邓勇进. 基于构件软件的可靠性通用模型[J]. 软件学报，2004，15(1)：27-32

[9] 单锦来，陈博，杨献春，等. MPEG-7和MPEG-7实验模型参考软件[J]. 计算机科学，2003，30(6)：31-37