术语自动提取中的领域度计算方法研究

张秦龙[1]，穗志方[2]，丁万松[3]

（1.北京大学计算语言所，1.北京，1.100871；2.北京大学计算语言所，2.北京，2.100871；3.北京大学计算语言所，3.北京，3.100871；）

摘　要： 术语自动提取是自然语言处理的主要内容之一，主要包括：单元度的计算和领域度的计算两方面。其中，领域度计算是术语自动提取区别于一般的新词发现等的关键步骤之一。本文详细阐述了术语自动提取中领域度的计算方法。通过综合利用领域部件信息和领域语料库的分类信息，借助机器学习方法探索了领域度的计算方法。选取计算机领域语料进行实验，并对实验结果进行了分析。实验结果表明，增加领域度信息后可以在很大程度上提高术语提取的准确率。

关键词： 术语提取；领域度；领域部件；SVM

The calculation of the Termhood in Automatic Term Extraction

Zhang Qinlong[1], Sui Zhifang[2], Ding Wansong[3]

(1-3.Institute of Computational Linguistics, Peking University 3.Peking 3.100871;)

Abstract: The Automatic Term Extraction is one of the most important contents in the NLP. It mainly contains the calculation of the Unithood and the calculation of the Termhood. The latter one is the key process which is different from the one in the neology extraction. In this paper, they raised the method of the calculation of the Termhood in the Automatic Term Extraction. They used the resources which is the lexical components of the specific domain and the taxonomic information coming from the domain corpus. To achieve the calculation of the Termhood, they made use of the Machine learning to handle these resources. Also, they chose the corpus based on the Computer domain to do several experiments. The results show that the improvement of the accuracy is prominent by adding the calculation of the Termhood.

Key Words: Term Extraction; Termhood; Domain Component; SVM

1　术语提取综述

术语是指在特定专业领域中一般概念的指称[1]（参见 GB/T 15237.1-2000 中华人民共和国国家标准 术语工作 词汇）。术语首先必须作为一个完整的语言单位出现，它必须具有出现频繁、结合紧密和使用自由的特点。其次术语作为专业领域中的一般概念，本身还应该有很强的领域性。

术语提取的主要任务就是通过综合考察术语的上述特征，从待处理语料中提取出术语来。术语提取是自然语言处理的主要内容之一，在信息检索、信息提取、数据挖掘等领域都有广泛的应用。目前术语提取的研究主要可以分为两个步骤：第一是判断一个符号串是否一个完整的语言单位；第二是判断这个语言单位是否特定领域的一般概念，即是否术语。我们可以将第一个步骤称为单元度的计算，第二个步骤称为领域度的计算。

目前的研究大部分都集中在单元度的计算方面[2][3]。一般是通过统计或者规则的方法来实现单元度的计

基金资助：本文研究得到东芝（中国）研究开发中心的资助

作者简介：张秦龙（1982-）男，陕西合阳，硕士在读，zql@pku.edu.cn

算。术语提取的第一阶段单元度计算的相关研究已经比较成熟，无论是在算法的效果还是效率等方面，都已经取得了较为令人满意的结果。然而单元度是从语言完整性的角度来判定一个字符串是否完整的语言单位，还不能作为衡量是否术语的唯一指标。在领域语料中，一个完整的语言单位不一定是一个领域术语。因此在满足单元度的基础上，我们需要从领域度的角度进一步考察。

与术语领域度计算相关的工作主要有利用信息检索领域广泛采用的 TfIDf 方法以及香港城市大学揭春雨博士提出的 rank 相减方法等。总体来说，在术语的领域度计算方面，目前还缺乏系统深入的研究。本文在 IT 领域分类语料库的基础上，通过综合利用领域部件信息和领域语料库的分类信息，借助机器学习方法探索了领域度的计算方法。实验结果表明，增加领域度信息后可以在很大程度上提高术语提取的准确率。

2 领域度计算的整体研究框架

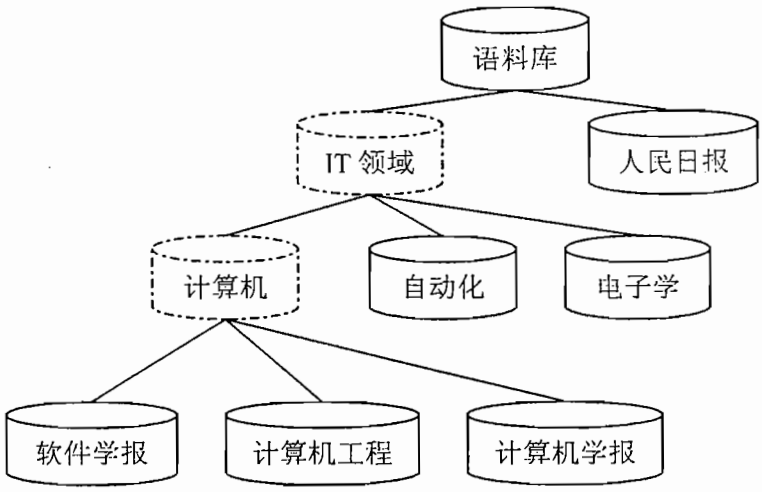

图 1 分类语料库

Fig.1 taxonomic corpus

本文的研究任务是针对 IT 领域的子领域计算机领域进行术语提取。根据现有的分类语料库资源（如图 1），在计算特定子领域术语领域度时，本文综合考察词语串在计算机领域中不同语料集合的出现，以及与计算机同级别的自动化和电子学等领域中的出现，以人民日报语料作为背景语料，来计算词语的领域度。同时根据词语串构成部件的领域性信息，来进行词语串领域性的衡量，最终得到语料中术语候选的列表。本文首先对语料文本进行自动分词，再使用 Nagao 串频统计方法进行半无限长子串的频率统计，在此基础之上使用谌贻荣[4]的单元度计算方法进行单元度的计算，得到完整的语言单位。然后进行领域度的计算，最终得到待处理语料的术语候选列表。系统整体框架如下：

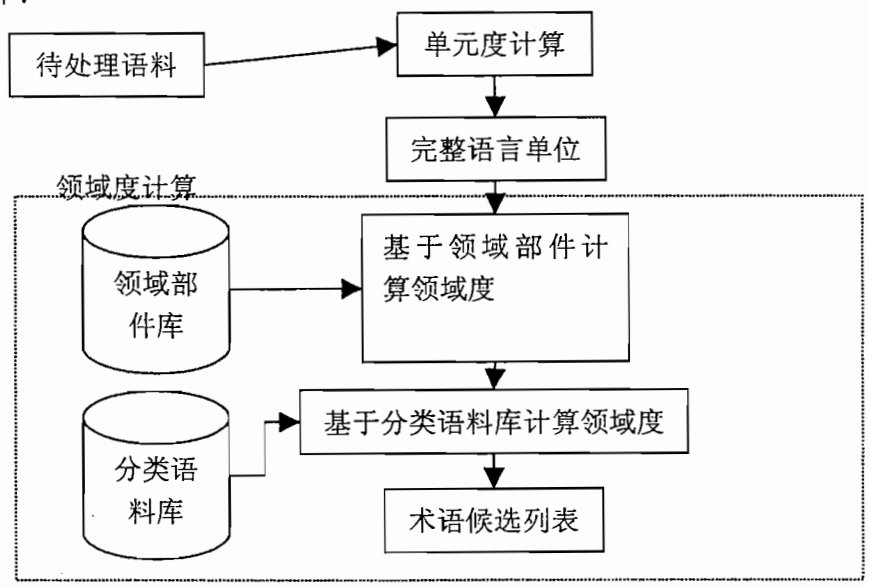

图 2 系统框架

Fig.2 System framework

3 基于领域部件计算领域度

领域部件指的是可以作为领域术语组成中具有领域性的成分。本文从分词、词性标注的 10 万 IT 领域中英文对照术语表（以下简称 IT 术语表）中提取出领域部件及其使用的位置及概率信息，来作为基于部件评价候选术语领域性的依据。

在提取领域部件时，本文主要利用了部件在术语内部出现的位置信息。首先通过对 IT 术语表的统计，生成一张包含领域部件位置及出现概率信息的统计表。统计表中包含的信息有，构件内容、在术语首位置出现概率、在术语中间位置出现概率和在术语末位置出现概率。不同位置出现的部件赋予不同的权值。同时在生成统计表的时候进行了平滑处理。具体计算方法[5]如下：

令部件内容为 W，IT 术语表中在术语首位置出现次数为 nWPrefix，中间位置出现次数为 nWMid，末位置出现次数为 nWSuffix，nW 为部件在 IT 术语对照表中出现的总次数，则：

$$P(WPrefix)=\frac{0.8\times nWPrefix+0.1\times nWMid+0.1\times nWSuffix}{nW} \quad (1)$$

$$P(WMid)=\frac{0.1\times nWPrefix+0.8\times nWMid+0.1\times nWSuffix}{nW} \quad (2)$$

$$P(WSuffix)=\frac{0.1\times nWPrefix+0.1\times nWMid+0.8\times nWSuffix}{nW} \quad (3)$$

通过上述计算，可以由 IT 术语表得到一张记录领域部件出现概率信息的统计信息表（如表 1、表 2、表 3 所示），作为使用领域部件判定术语的依据。

基于部件的领域度计算公式如下：

$$termhood(D)=D.pPrefix\times(D.pMid1\times D.pMid2\times..)\times D.pSuffix \quad (4)$$

其中 D 为候选术语，D.pPrefix 为在 D 首位置出现的领域部件的概率，D.pMidi(i=1,2,..)为在 D 中中间位置出现的部件的概率，D.pSuffix 为 D 中末位置出现的部件的概率。由于我们的候选术语保留了切分信息，因此对于候选术语的每一个切分单位，分别查询其在统计信息表中相应概率值，然后将各部分查得的概率值相乘，即得到候选术语的领域度。

4 基于分类语料库计算领域度

本文在图 1 所示的语料库中计算计算机领域的领域性。如果令集合 A：计算机领域候选术语；B：与计算机同层次的电子、自动化等领域语料；C：人民日报语料；则术语在其中的出现情况如图 3 所示。

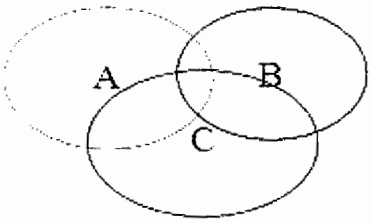

图 3 领域分布

Fig.3 Domain distribution

4．1 基于 SVM 的方法

特定子领域的候选术语领域性计算问题，可以看作是对待处理语料中的候选术语进行分类的问题。我们可以以特征的观点来看待集合之间的作用，使用机器学习中的 SVM 方法来进行术语的提取。

SVM 的基本思想上把特征看作多维空间中的点，通过训练得到一个最优的曲面，使得这个曲面尽可能的把空

间中的那些点划分成两个部分。如图 4 所示，以两维的情况为例，实心点和空心点代表不同的两类样本。H 为期望通过训练得到的分类面，H1 和 H2 是平行于 H 且过离分类面最近的样本。SVM 训练的目的就是使得 H 能够正确的将两类样本分开，且使得 H1 和 H2 之间的距离尽可能的大。

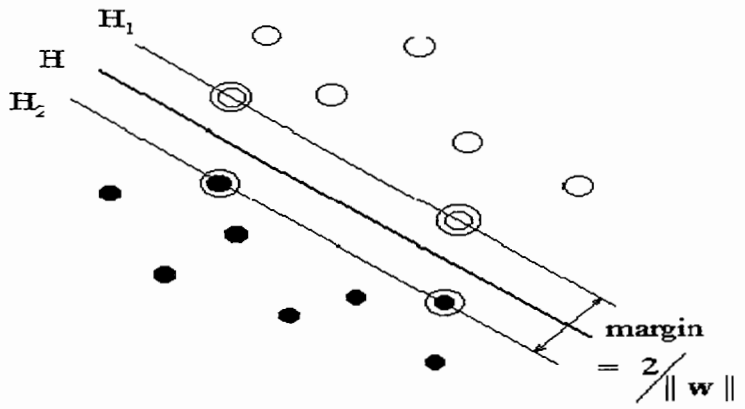

图 4 最优分类面

Fig.4. Optimum hyperplane

本文将 A 类语料看作一类样本，B 类和 C 类看作另外一类样本。对候选术语的提取可以看作是将候选术语分类到 A 类和 B，C 类中的过程。我们可以利用 SVM 的方法来实现这个目标。本文选取三类特征作为分类特征：第一类特征是待处理词串在 A 类领域语料出现的概率；第二类特征是待处理词串在 B 类领域语料出现的频率；第三类特征是待处理词串在 C 类领域语料中出现的频率。使用 SVM light 工具进行训练和分类，取得了较好的效果。

5 实验

5．1 实验语料

实验使用的语料是《计算机工程》1999 年 4 月到 1999 年 12 月（不包括 6 月）语料（以下简称《计算机工程》）、《软件学报》1998 年 1 月到 2000 年 4 月语料（以下简称《软件学报》）、《电子学报》1998 年 1 月到 1998 年 5 月语料（以下简称《电子学报》）和《人民日报》1998 年 1 月语料（以下简称《人民日报》）。所有语料首先进行自动切词，并使用 Nagao 串频统计算法进行串频统计。

5．2 领域部件库的建立

利用领域构件计算领域度。首先对人工标注的 IT 术语表进行统计，得到标注术语首位置、末位置以及中间位置出现的部件频率信息，然后应用公式（1）、（2）和（3），计算得到领域部件的统计信息表。举例如下：

表 1 首位置部件

Tab.1 Occurring in the first place

术语部件	首位置出现概率
大面积	0.8000000000000002
单程	0.8000000000000002
多机	0.8000000000000002
再生	0.4911764705882353

表 2 末位置部件

Tab.2 Occurring in the last place

术语部件	末位出现概率
编码法	0.8
测量法	0.8
定法	0.8
估计法	0.8

表 3 中间位置部件

Tab.3 Occurring in the middle place

术语部件	中间位置出现概率
半导体场	0.8000000000000002
电联	0.8
超高频	0.45
良好	0.45
数据	0.4466557911908645?
分量	0.4362204724409449
中继	0.4402777777777777

5．3 评测原则

为了进行比较，本文分别就单纯单元度计算、领域部件计算领域度、SVM 方法计算领域度、领域部件综合 SVM 方法计算领域度等进行了实验。实验使用准确率和召回率来对结果进行评价。准确率的计算以人工标注的 IT 术语表为基准，如果待评价词串是 IT 术语表中的术语的父串且只比 IT 术语表中术语多一个语言单位，或者待评价词串和 IT 术语表中的术语只相差一个语言单位，则认为提取正确。例如 IT 术语表中只收录了“正弦曲线”，但是在评价待选词串“余弦曲线”时，我们也认为“余弦曲线”是一个术语。

$$准确率 = \frac{提取出的部分匹配术语个数}{语料中提取出的术语个数} \tag{5}$$

$$召回率 = \frac{提取出的术语个数}{语料中出现的术语个数} \tag{6}$$

5．4 实验结果

实验一：单纯单元度计算。在本实验中仅应用单元度计算方法来进行术语提取。

实验二：基于领域部件计算领域度。

实验三：应用 SVM 方法计算领域度。我们采用《计算机工程》语料作为训练语料，《电子学报》和《人民日报》语料分别作为 B 类领域和 C 类领域语料。使用 SVM light 工具包进行训练。

实验四：综合领域部件和 SVM 方法计算领域度。在实验二处理的基础之上，应用 SVM 的方法进行计算。得到结果如图 5 和 6 所示：

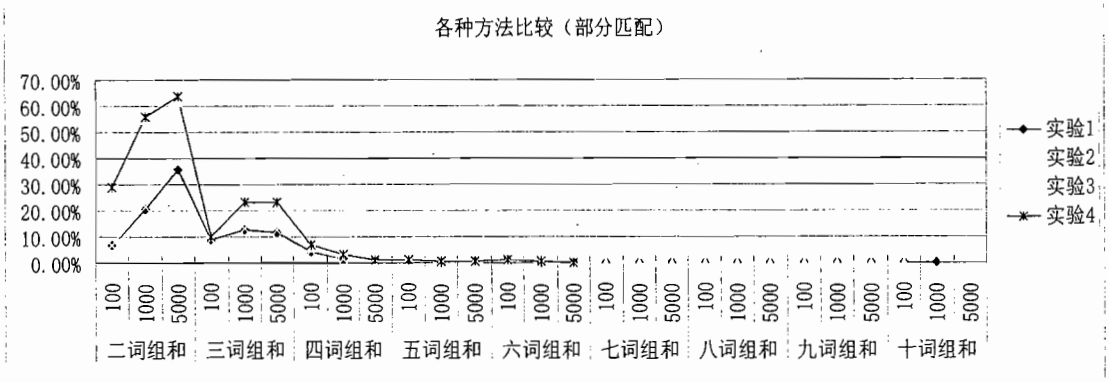

图 5 实验结果比较一

Fig.5. Result I

由以上数据可以看出，二词组合的术语提取中，单纯使用单元度进行术语提取（实验 1）的正确率在 20.60%（前 1000 个）和 35.76%（前 5000 个），而增加领域度计算之后（实验 4），术语自动提取的正确率可以提高到 55.80%（前 1000 个）和 63.54%（前 5000 个）；即使是在三词组和的术语提取中，单纯使用单元度计算正确率只有 12.40%和 11.62%，而增加领域度计算之后的正确率为 23.00%和 23.06%，增加领域度计算的方法也明显好于单纯的使用单元度来进行术语提取的结果。说明了领域度计算对于术语自动提取是非常重要的。

5．5 实验结果分析

将上述实验得到的结果曲线叠加，我们可以发现使用领域部件的方法对提取效果影响最大，这说明领域部件在判定术语的过程中起了至关重要的作用。可见在术语提取中，充分考虑术语的特定用词及特征是非常重要的。

而 SVM 方法的结果不尽如人意，这可能主要以下两方面的原因：（1）特征的选择问题：目前特征的选择是以该词串是否在 A 类或 B 类或 C 类语料中的出现频率或概率为依据的。而当提取多词组合的术语时，由于每个候选词串出现的频率一般都比较低，因此在此情况下，频率可能不能更好的刻画术语的特征。（2）数据稀疏的问题：当待处理语料中词语在其他类别中很少出现导致数据稀疏时，将直接影响 SVM 的训练效果。

由于本文术语提取的最终结果是以 N 词词语组合术语候选列表的形式呈现。因此无法详细统计 N 取不同值时

的召回率。最后通过公式6，可得到整个提取结果的召回率为57.14%。

6 进一步的工作

总的来说本文在术语提取领域度计算方面提出的两个方法是可行的，并且取得了较好的效果。但是还存在一些不足，下一步可能需要在SVM方法的特征选择方面仔细考虑，以提高提取的准确率。另外一个就是评测方法还有不尽人意的地方，如何制定一个合理的，且能真正反映术语提取效果的评测方法，可能也是下一步继续努力的方向。

参考文献：

[1]. 中华人民共和国国家标准 GB/T 15237.1-2000 《术语工作 词汇》

[2]. 罗盛芬，孙茂松. 基于字串内部结合紧密度的汉语自动抽词实验研究[J]. 中文信息学报 2003 第17卷 第3期:pp.9-14

[3]. 谌贻荣. 内部紧密度和边缘自由度相结合的符号串单元率估算. 全国第八届计算语言学联合学术会议（JSCL-2005）

[4]. 谌贻荣. 中文术语自动提取技术研究[D]. 保存地点：北京大学图书馆，2005

[5]. 穗志方. 自然语言处理技术在术语标准化工作中的应用[A]. 徐波,孙茂松,靳光瑾. 中文信息处理若干重要问题 科学出版社 2003.11