第 33 卷 第 2 期 计 算 机 工 程 2007 年 1 月
Vol.33 No.2 Computer Engineering January 2007
·软件技术与数据库· 文章编号：1000—3428(2007)02—0047—03 文献标识码：A 中图分类号：TP311

自动提取含字母词语的领域新术语的研究

姜韶华，党延忠
（大连理工大学系统工程研究所，大连 116024）

摘 要：新术语的提取是中文信息处理领域的一个重要研究课题。针对现有提取方法的不足和很多专业术语表现为字母词语的特点，该文提出了一种综合统计技术和规则筛选的方法：基于长串优先和串频统计的思路进行文本切分，得到共现字符串，利用词语搭配规则进行过滤，经过领域词典及评价函数的筛选，提取出领域新术语。该方法可发现包含字母词语、专业术语等未登录词在内的频率大于等于 2 的任意长度的专指语义串、短语和词。实验表明了该方法的有效性及新术语的准确率分布特征。

关键词：专指语义串；长串优先；字母词语；中文信息处理

Research on Automatic Extraction of Chinese New Domain-specific Terms Comprising Lettered-words

JIANG Shaohua, DANG Yanzhong
(Institute of Systems Engineering, Dalian University of Technology, Dalian 116024)

【Abstract】Extraction of new domain-specific terms is one of the important topics in Chinese natural language processing. Aiming at the limitation of the current methods and the specialties of many domain-specific terms are lettered-words, a novel approach combined with statistic technique and rule is proposed to extract new special semantic strings. Co-occurrence of character strings is formed by text segmentation based on matching longer strings first combined with frequency statistics. No-meaningful character strings are trimmed by collocation rules. Filtered by domain lexicon and membership degree, new domain-specific terms are extracted finally. This method can extract new special semantic strings, phrases and words, including unknown words like lettered-words and domain-specific terms, their frequency is larger than 2. Experiments show that this extraction technique is effective and indicate new domain-specific terms' distribution characteristic of precision ratio.

【Key words】Special semantic strings; Matching longer string first; Lettered-words; Chinese natural language processing

1 概述

科学技术的快速发展不断产生新的术语和专有名词等未登录词，其中相当部分首先以字母词语的形式出现而得到广泛应用[1]，目前关于自动提取字母词语的工作还不多见。包括术语、专有名词、字母词语的领域词汇集中体现一个学科领域的概念和知识，有助于掌握领域的现状和趋势。随着中文信息处理应用领域的不断扩展，对于发现领域术语的需求也越来越迫切[2]。人工提取效率低下、难以更新，因而迫切需要能够自动提取术语的方法。

目前主要有两个提取术语的方法：基于规则方法和基于统计的方法。基于规则的方法，其核心是根据语言学原理和知识制定一系列共性规则和个性规则以处理自动分析中遇到的各种语言现象。但自然语言难以用一套规则去准确地预测真实文本中所出现的各种现象。随着互联网的快速发展，可以便捷地获取机器可读的大量语料，基于统计的方法成为目前研究的主流。现有的统计方法[3]往往从两字词开始扩展到多字词，统计量过大。

本文提出了长串优先的统计技术与规则优化相结合的新术语提取方法。依据最少分词原则[4]，首先将待处理的语料采用包括切分符号和切分汉字串的切分标记进行预处理；其次采用长度优先与串频统计的方法切分出频率大于等于 2 的所有字符串；再采用 3 种词语搭配规则过滤不成词的共现字符串；然后与领域词典作对比分析；最后根据评价函数筛选，最终提取出包含专指语义串及字母词语的新术语。上述思路可表示为如图 1 所示的流程。

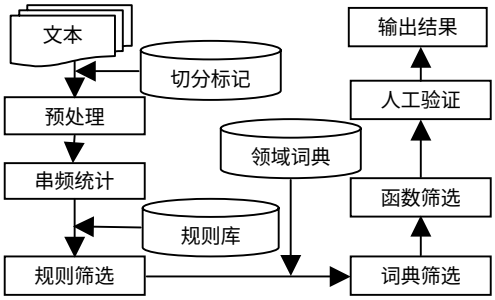

图 1 新术语抽取流程

2 基本概念

定义 1 汉字串，是文本中连续的汉字组成的字符串。

定义 2 字母词语，是指由拉丁字母（包括汉语拼音字母）或者希腊字母构成的或由它们分别与汉字混合构成的词。

定义 3 专指语义串，是指由短语、词或短语和词的组合构成，在语义上比短语更具有专指性的字符串。如“管理信息系统”，若用词典切分，则结果为“管理”、“信息”、“系统”，它们的组合并不能准确表达出“管理信息系统”作为一个整体所具有的语义，因而专指语义串更全面地体现了概念的整

基金项目：国家自然科学基金资助项目(70271046)
作者简介：姜韶华(1971－)，男，博士生，主研方向：数据挖掘，知识工程，系统开发与集成；党延忠，教授、博导
收稿日期：2006-01-24 **E-mail：**shjiang@xinhuanet.com

体性。

定义 4 切分符号，是指领域术语中不可能出现的符号。包括自然切分符号与非自然切分符号两类。自然切分符号指标点符号，非自然切分符号包括数字等不能构词的符号。

定义 5 切分汉字串，是指不能构词的单字和独立使用且具有确定意义的汉字串。包含出现频率极高的单字虚词，如的、了、又、与、和、及、能等；常用的双字虚词：如基于、以及、关于、对于、用于等；经常出现的实词，对于科研性质的文本来说，主要包括通知、报告、研究、实现等。

定义 6 共现字符串，是文本中出现频率大于 1 的固定搭配的字符串。

定义 7 候选术语集（记为 Ω），是存放作为候选术语的共现字符串的集合，其结构描述为：<候选术语集> :: = {<共现字符串>, <串频>}。

3 提取方法

3.1 长度优先的统计模型

书面汉语的词汇形态十分稳定，因此连续的字符串在上下文中共现的频率越高，则构词的可能性越高。根据术语平均词长较长的特点，依据长字符串优先、串长度递减的切分原则，统计得到具有频率信息的共现字符串集合。算法描述如下：

maxlen 为欲提取的最大字符串长度，minlen 为欲提取的最小字符串长度

```
while (maxlen > minlen)
 {
        startpos =  字符串起始位置
        endpos =  字符串后空格位置
        tempstr = endpos - startpos
        if (tempstr 的长度小于 maxlen)
        {进入下一个字符串}
        else {
              tempstr = startpos + maxlen
              if (文本中不能匹配 tempstr)
              {从下一个字符开始继续提取 tempstr}
              else {
Ω = Ω ∪ tempstr（包括串频）
}
        startpos = endpos + 1
        endpos = startpos 后空格位置
        }
 maxlen = maxlen – 1       }
```

汉语是大字符集的语言。一个汉字为两个字节，而西文为单字节。由于绝大多数单个汉字不具备独立的信息量，本文设置欲提取的最小字符串长度 minlen 大于 2。

串频最大匹配法的优点是从最长的共现字符串开始匹配，这样就避免了长字符串中包含的子字符串的无效匹配，并且可以识别出候选字母词语和专指语义串。

3.2 规则优化

由于统计的方法没有考虑词法、语法、语义信息，因此候选术语集 Ω 中可能存在错误的共现字符串。通过对错误共现字符串的观察研究，提出如下 3 条规则。通过这 3 条规则的筛选来提高候选术语集的准确率。

规则 1 删除不合理的介词搭配

实验发现大部分错误是由于一些介词，如“在”、“中”、“向”等导致噪声，因此需要重点加以筛选。将介词记为 c_i，包含 c_i 的固有词条记为 t_{ij}。筛选过程如下：

对候选术语集 Ω 中的每个共现字符串
IF（(共现字符串包含 c_i）AND（共现字符串不包含 t_{ij}））
THEN（从该共现字符串中删除 c_i）

规则 2 删除“数词＋量词”的共现字符串

“数词＋量词”的共现字符串，如“一个”等，往往不是领域词，因此需要删除。

规则 3 删除不符合词首字规律的共现字符串

有些词不能出现在词首，如“髓、殄、茎、鳍”等，有些词只能出现在词首，如“篇、挨、催、粘”等。

经过规则筛选后，将相同共现字符串的频率累加，如下所示：

IF(Ω 中有相同的共现字符串)
THEN（合并相同的共现字符串，将对应频度求和）

3.3 领域词典过滤

本文以历史领域词集合作为领域词典，新术语应该在领域词典中没有出现过。在获得的候选术语集 Ω 中，通过与领域词典的对比分析，从而发现并提取出新的领域术语。

3.4 评价函数筛选

新术语，是一种固定的字符串组合。其特点是不仅要在一篇文章中多次出现，而且要在多篇文档中反复出现[5]。

用 D 表示领域词典，其中的元素记为 d，$D=\{d|d=(string)\}$。S 表示经过规则优化和 D 过滤的共现字符串集合，其中的元素记为 s，$S=\{s|s=(string,totalfreq,textfreq)\}$。其中 *string* 表示字符串，*totalfreq* 表示总的串频，*textfreq* 表示文本频率，即该串出现的文本篇数。则评价函数可表示如下：

$$f(s)=\begin{cases}1, if\ (s.totalfreq \geq M\ \&\ s.textfreq \geq N\ \& \\ \forall s, s.string \neq d.string\ ) \\ 0, other\end{cases}$$

其中，M 和 N 为阈值。函数值为 1 即表示共现字符串为新术语。

在函数值为 1 的共现字符串基础上，经过人工验证可得到最终的判定结果。

4 实验及分析

4.1 实验数据

为测试新术语提取的效果，选取信息及电子领域从 1999 年到 2004 年的学术文章的标题、关键词和摘要作为实验语料，并将 1999~2003 年的术语集合作为领域词典，在 2004 年的文本上作新术语提取研究。具体语料情况如表 1 所示。

表 1 实验语料

年份	篇数	大小(KB)
1999~2003	3 182	1 319
2004	1 320	698

4.2 实验结果

根据新术语的特点及本文采用语料情况，取 m=4，n=2。

提取新术语 1 053 个，其中字母词语 93 个。准确率指标衡量的提取结果如表 2 所示。

准确率＝人工判定正确的新术语数/提取的新术语总数×100%

表 2 准确率结果

平均准确率	字母词语准确率	非字母词语准确率
94.3	87.1	95.0

将新术语按频率从低到高排序，分别计算不同频率的新术语的累计准确率，由于频率大于 12 的错误术语很少，因此合并到频率为 12 中一起分析。新术语的累计准确率分布趋势

如图 1 所示。

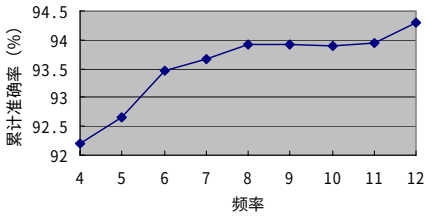

图 1 不同频率累计准确率分布

为考察非字母词语中错误术语的分布与术语长度的关系，分别统计非字母词语中不同字长错误术语占总错误数量的比例，具体如图 2 所示。

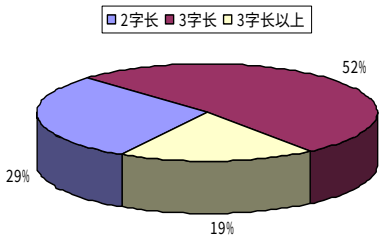

图 2 不同字长错误术语分布比例

下面是实验提取的部分新术语：

二字新术语，如蠕虫、工况、陀螺、异步、短信、隐写、光频、射流、对偶、粗集…

三字新术语，如本体论、供应链、潜信道、多相流、绝缘子、螺旋藻、转基因、磁控管、...

四字新术语，如北斗卫星、异向介质、网格计算、网络编码、蚁群算法…

四字以上新术语，如小世界网络、虚拟骨干网、自适应算法、磁电阻抗效应、二代小波变换…

字母词语，获至宝如 QoS、Wavelet、BP 网络、Petri 网、多 Agent 协作、多层螺旋 CT…

4.3 实验数据分析

从准确率实验结果看出，本文提出的方法提取新术语的准确率比较高，平均可达 94.3%。字母词语的准确率较非字母词语低，这主要是由于字母词语数量较少，而将部分字母词语的公共部分提取出来作为结果而引起的，如“ company ”和“ compile ”的提取结果为“ comp ”。

从不同频率累计准确率分布图可知，准确率随着术语频率的增大而增加，这意味着选取高频术语会获得更好的结果。

不同字长错误术语分布比例的实验结果表明 3 个字长的术语错误数量比例最高，超过一半，达到了 52%，而 2 个字长的术语次之，为 29%，3 个字长以上的术语错误数量比例仅占 19%，因此优先提取 3 个字长以上的术语会取得更好的效果。

5 结束语

新术语的不断出现是一个客观现象，而互联网的发展可以提供丰富的语料资源。本文提出的基于长度优先和串频统计思想的方法，综合统计技术和规则优化的优点，通过领域词典的过滤和评价函数的筛选，最终提取出包括专指语义串和字母词语等未登录词在内的新术语。本文的工作将有助于提高中文分词、信息检索、搜索引擎、机器翻译等的质量。下一步的工作是如何更好地完善规则库和提高字母词语的准确率。

参考文献

1 郑泽之，张　普，杨建国. 基于语料库的字母词语自动提取研究[J]. 中文信息学报, 2005, 19(2): 78-85.

2 孙　霞，郑庆华，王朝静，等. 一种基于生语料的领域词典生成方法[J]. 小型微型计算机系统, 2005, 26(6): 1088-1092.

3 Pantel P, Lin D. A Statistical Corpus-based Term Extractor[M]. Lecture Notes in Artificial Intelligence. Springer-Verlag, 2001: 36-46.

4 刘秉权，王晓龙，王宇颖. 一种多知识源汉语语言模型的研究与实现[J]. 计算机研究与发展, 2002, 39 (2): 231-235.

5 邹　纲，刘　洋，刘　群，等. 面向 Internet 的中文新词语检测[J]. 中文信息学报, 2004, 18(6): 1-9.

(上接第 36 页)

和方法也是基于此原因而提出的，比较而言，减少数据传输耗时可通过搭建高效数据网格及提交网络通信环境来解决。

5 总结

GIS 模型实现网格计算主要有 3 种加工方法：模型分块加工、分步加工和立体加工法。模型计算可按分布式计算特性划分为串行、并行和串并行混合计算模式。本文主要列举了 GIS 应用几种常用模型的并行算法，有直方图、空间数据内插。在 GIS 应用中，这类算法具代表性、可扩展性及基础性等特点，可作为同类网格并行算法的扩展基础。这些算法的实现原理是：数据均衡分块，每个网格节点机处理分块数据，利用相应空间分析算法进行并行计算。

原理易实现，可操作性强，因此，算法实例都是基于原理给出的。本文提出矢量地图数据在网格计算之前的切分规则：最大面积法和结点最多法，目的是增强数据并行计算之前的数据分割及之后的整合效率，提高并行计算效果。

参考文献

1 肖辉力，李　京，陈秀万，等. 地理信息系统的模型库研究[J]. 地学前缘, 2000, 7(增刊).

2 Foster I, Kesselman C. The Grid: Blueprint for a New Computing Infrastructure[M]. Morgan Kaufmann, 1998: 11-16.

3 蒋艳凰，杨学军，易会战. 卫星遥感图像并行几何校正算法研究[J]. 计算机学报, 2004, 27(7).

4 Wang Shaowen, Armstrong M P. A Quadtree Approach to Domain Decomposition for Spatial Interpolation in Grid Computing Environments[J]. Parallel Computing, 2003, 29(10): 1481-1504.

5 Erik G, Samet H H. Data-parallel Polygonization[J]. Parallel Computing, 2003, 29(10): 1381-1401.

6 First Results and Future Perspectives of the European Data Grid Project[EB/OL]. http://www.hoise.com/primeur/02/articles/weekly/AE-PR-04-02-22.html.

7 David A B. A Framework for the Integration of Geographical Information Systems and Model Base Management[J]. Geographical Information Science, 1997, 11(4).