词干提取方法及工具的对比分析研究*

吴思竹 钱 庆 胡铁军 李丹亚 李军莲 洪 娜
中国医学科学院医学信息研究所 北京 100020

〔摘要〕首先，论述词干提取和语言形态学的关系。其次，对词干提取的主要算法和相关技术进行详细分类。而后，对词干提取方法的主要特点进行分析并对各类方法的优点、缺点和特点进行对比，并归纳实现词干提取功能的工具及其特点。最后，对词干提取研究的现有问题和未来发展趋势进行探讨。

〔关键词〕词形规范化 词干提取 波特算法 语言形态学
〔分类号〕G250

Comparative Analysis of Methods and Tools for Word Stemming

Wu Sizhu Qian Qing Hu Tiejun Li Danya Li Junlian Hong Na
Institute of Medical Information of Chinese Academy of Medical Sciences, Beijing 100020

〔**Abstract**〕The authors talk about current stemming methods and tools. Firstly, they discuss the relationship between stemming and morphology. Secondly, they describe the categories about stemming algorithms in detail. Thirdly, they summarize the characteristics of various methods and tools, and compare their strengths and weaknesses. Finally, they discuss the problems and development of the research on stemming.

〔**Keywords**〕word normalization stemming Porter stemmer morphology

词形规范化处理是信息检索系统及文本处理过程中的一种必要的基础操作，它为一组相关的词形识别出一个具有公认性的、标准化形式的代表。词干提取是词形规范化处理的重要技术之一，主要应用于信息检索和文本处理。在检索系统中，对文本中的词进行词干提取，能够减少词的数量，缩减索引文件所占空间，并且使检索不受输入检索词的特定词形的限制，扩展检索结果，提高查全率。本文主要对词干提取方法和工具进行归纳和分类，对主要方法及工具的特点进行对比分析。

1 词干提取和语言形态学

词干提取是词形规范化处理的重要技术之一，将一个词的不同形式统一为一种具有代表性的标准形式（词干）。它主要处理的对象为语言中的词，需要了解语言词汇的结构，因此与语言形态学（morphology）密不可分。语言形态学是一个关注于词汇内在结构和构词规则的语言领域，研究如何从比较小的意义单位（语素）构成词。语素是语言中负荷意义的最小单元，包括词干（stem）和词缀（affix）两大类。词干承担词的主要意义，而词缀提供附加涵义。语素构成单词的方法可以分为三大类（可能有部分的交叉）：屈折（inflection）、派生（derivation）和复合（compounding）。

屈折和派生均是把词干和一个语法语素结合起来。屈折是相对而言较为简单的语言变化形态，词在屈折变化前后，词性基本不发生改变，主要包括名词复数、名词所有格和动词过去式、过去分词、现在进行时等[1]，变化规律性较强。如英语名词复数变化，一般添加后缀为“s”、“es”，动词过去式添加为“ed”，动词现在进行时添加“ing”等。派生相对复杂，变化前后形成的单词一般属于不同的词性，例如，动词 computerize 可以加上派生后缀“ation”，形成名词 computerization；并且，派生形态产生的词，其涵义可能会发生变化。复合是两个或两个以上独立的词在一起组成新词，如 dog catcher。复合形态更为复杂，可能包含屈折或派生的

* 本文系国家“十二五”科技支撑计划项目“科技知识组织体系的协同工作系统和辅助工具开发”（项目编号：2011BAH10B02）和中国医学科学院医学信息研究所基本科研业务费课题“基于语言网络的医学文本表示模型构建方法研究”（项目编号：11R0209）研究成果之一。
收稿日期：2012－01－11 本文起止页码：109－115，142 本文责任编辑：王善军

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

变化。

词干提取主要就是根据语言形态中的规律进行处理的，去除屈折或派生形态的词缀，获得词干。因此，只有掌握语言构成特点，深入解析语言形态变化，才能发现其中规律，提高词干提取的准确性。目前，在信息检索和文本处理应用当中，词干提取还是较为浅层的词形规范化技术，不考虑词性、语义等复杂问题，主要是进行词形的统一。

2 主要方法分类

词干提取方法的发展已较为成熟，有很多经典算法和应用。

2.1 词干提取方法分类

本文按照各方法的实现原理，主要将词干提取的方法归纳为基于规则的方法、词典查找法、基于统计的方法和混合的方法 4 类，并根据各类方法又进行更细致的划分，具体划分如图 1 所示：

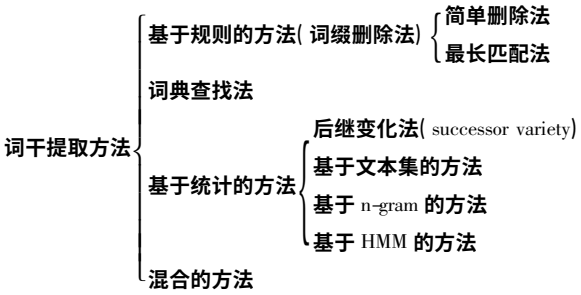

图 1 词干提取方法主要类型划分

2.1.1 基于规则的方法 基于规则的方法主要是识别语言形态构成中的特有规律，进行词干提取或词缀消减的方法。一般而言，词缀包括前缀、中缀、后缀，在基于规则的词干提取方法中，主要是通过形态规则去除词的后缀。规则主要基于特定语种的语言形态变化如英语中屈折形态变化，名词单数到复数形式的转化规律有：绝大多数可数名词的复数是在该词结尾加上后缀“s”；以“s”、“z”、“x”、“ch”、“sh”结尾的词加上后缀“es”；以辅音 + “y”结尾的，改“y”为“i”加后缀“es”等。词干提取利用这些规则进行词缀的删除，还原词干。该类方法又可细分为简单删除法和最长匹配法。简单删除法主要处理名词复数的后缀，典型的如 S - stem 算法，该种算法简单、快捷，但精确度低，无法处理更多的词汇变形形式。另一类是最长匹配算法，主要是将词和基于规则变化的后缀进行匹配，删除最长匹配的后缀，其余部分作为词干。该类方法包含很多经典算法，获得广泛应用。经典的算法主要有 Porter 算法[2]、Lancaster (Paice/Husk) 算法[3]、Lovin 算法[4] 和 Dawson 算法[5]。

——Porter 算法。Porter 算法是一个用于删除英语单词中通用形态和屈折词缀的算法，由剑桥大学的 Martin Porter 于 1980 年开发出来，被广泛应用于信息检索系统中，如 Lucene、Solr、Google 等，主要用于术语的标准化。在 Porter 算法中，任何词都可以表示为：

$$[C](VC)^m[V]$$

其中 C 为辅音，V 为元音，[] 表示内容可选项出现，m 被称为对任意词的测度，表示 VC 重复了 m 次。m 的值可以取任何大于等于 0 的值，用于决定是否需删除一个已知词缀。

算法规则的表现形式为：

$$(condition)\ S1 -> S2$$

它表示若满足条件，词缀 S1 则被替换为词缀 S2。

Porter 算法将词干提取划分为一系列线性步骤，主要有：①处理名词复数(删除词缀或对复数形式重新编码)，处理动词“ed”或“ing”结尾的单词，并对词干重新编码；②当词干中包含元音，并以“y”结尾时，将“y”改为“i”；③双词后缀结尾的词，将双词缀映射为单词缀，如将“ization”映射为“ize”等；④处理“ic”、“full”、“ness”等词缀；⑤在 [c] vcvc [v] 情形下，处理“ant”、“ence”等后缀；⑥删除最后的“e”。如图 2 所示：

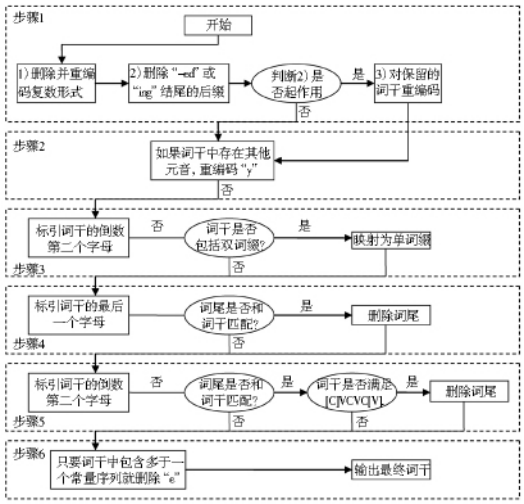

图 2 Porter 算法主要流程

——Lancaster (Paice/Husk) 算法。Lancaster 算法是一种简单的迭代算法，通过一些循环步骤删除词缀。它使用一个规则文件，处理时提取一个词的尾字母，使用索引查找该字母适用的规则。规则格式基本由 5 个部分组成：①逆序排列的词尾；②一个可选项的完整的

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

标志;③指定删除字符数;④可选附加字符串;⑤连字符“>”或“.”。如:规则“nois4j>”表示以“sion”结尾的词将“sion”这 4 个字符替代为“j”;规则“de2>”表示以“ed”这两个字符结尾的词去掉“ed”。

该算法具体实现的主要步骤如下:①检查词缀字母,是否和规则匹配。如不匹配,输出词干;匹配,则进行下一步操作。②检查规则的可用性。可用,生成新词干并进行下一步;否则访问下一条规则并返回步骤①。③检查新词干是否通过检查认定。是,则进行下一步;否则转换为旧的词干,访问下一条规则并返回步骤①。④判断是否还需再次提取词干。是,则返回步骤①;否,则输出最终词干。如图 3 所示:

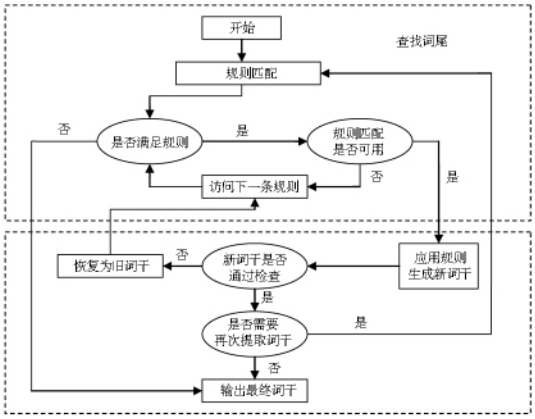

图 3 Lancaster (Paice/Husk) 算法主要流程

——Lovin 算法。Lovin 算法由 Julie Beth Lovins 于 1968 年首次提出。它基于最长匹配原理删除词的后缀,是一种单程、上下文敏感的词干提取算法。该算法主要应用一个包含了 294 个词后缀的列表,并结合了约 29 个约束条件和 35 个转换规则。按词后缀包含的字母数降序排列,每个词缀对应一个约束条件,共 29 个约束条件,编号从 A 到 Z,再加上 AA、BB、CC。这些条件对词干长度、什么情况下删除哪些词缀或不删除词缀等进行了约束。35 个转换规则主要是进行词干字符的替代转换,用于消除拼写错误,将“umpt”替换为“um”等,这也是一个对词干进行重编码的过程。Lovin 的词后缀列表如图 4 所示:

```
.11.
alistically B arizability A izationally B
.10.
antialness A arisations A arizations A entialness A
.09.
allically C antaneous A antiality A arisation A
arization A ationally B ativeness A eableness E
```

图 4 Lovin 的词后缀列表

Lovin 算法主要包括两个步骤:①词干提取阶段,包括词缀删除和相关规则的实施。首先,要决定从词缀列表中开始查找匹配的位置,并规定词干至少要保留两个字母长度。而后,进行词缀列表查找及匹配。如匹配失败,进行重新编码;如匹配成功,找到词缀,则判断是否满足规则。若满足,则删除词缀;不满足,则进行重新编码。②重新编码。初步获得词干后,查询转换规则列表,进行匹配。如匹配成功,则根据规则对词干重新编码,将词干转换为有意义的词,输出最终词干。如没有可用的规则,则直接输出最终词干。如图 5 所示:

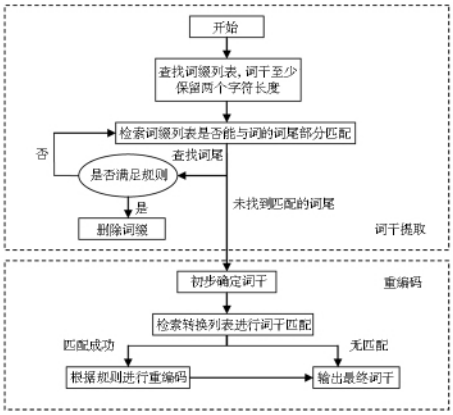

图 5 Lovin 算法主要流程

——Dawson 算法。Dawson 算法由 John Dawson 于 1974 年提出,主要是对 Lovin 算法的规则和匹配方法进行了改进,并解决了拼写异常问题。Dawson 算法采用部分匹配,尝试匹配在特定限制条件中相同的词干。其主要工作包括两部分:一是对所有复数和简单词缀进行了归并,增加了词后缀列表;二是利用词后缀列表和完整性匹配原则实现了所有词形变化的归并。

2.1.2 词典查找法 词典查找法也就是利用暴力算法(Brute Force)[6],主要通过查找已知词典进行词形的转换,词典中包含词的词干形式和对应的多种词形。已知一个词,查找词典,找到相同的词形,而后返回对应的词干。该类方法中具有代表性的是 Krovetz 方法。

Krovetz 算法[7]由 Robert Krovetz 于 1993 年提出,是一种基于语言词典有效性的词干提取方法。词典由几个列表构成,包括词的基础列表、补充附加词表、规则外异常词列表、直接异形合并列表、国家国籍列表、专有名词列表。算法首先进行后缀删除,而后进行词典检测,发现规则外的异常进行重新编码,最后返回词干。Krovetz 算法返回的词干与其他词干提取算法不同,其返回的词干为完整的词而不是删减后的不完整的部分词。Krovetz 算法相对于其他算法的优势在于算法易于修改,更加灵活,但是缺点是相对更加依赖于

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

词典的收录范围。

2.1.3　基于统计的方法　为了解决对词典未收录词无法进行词形规范的问题，研究者尝试通过词的分布统计规律进行词干的提取。现有研究可分为后继变化法、基于文本集的方法、n-gram 方法和 HMM 方法。

——后继变化法（successor variety）。后继变化法主要是由 Hafer 和 Weiss 提出的词干提取方法[8]。与词缀删除法不同，它主要是通过文本集中的字符分布统计，识别词干和后缀边界，直接提取词干。一个字符串的后继变化数是指在文本集的词中，跟在这个字符串后包含的不同字符数。例如，文本中的字符串为：ABLE，BEATABLE，FIXABLE，READ，READABLE，READING，RED，ROPE，RIPE。READABLE 的后继变化数如图 6 所示：

文本集
ABLE, APE
BEATABLE
FIXABLE
READ
READABLE
READING
READS
RED
ROPE, RIPE

前缀	后继变化数	字母
R	3	E,I,O
RE	2	A,D
REA	1	D
READ	3	A,I,S
READA	1	B
READAB	1	L
READABL	1	E
READABLE	1	BLANK

图 6　文本中 READABLE 的后继变化数统计[8]

显而易见，READ 的后继变化数为 3，为最后词干。

该算法进行词干提取的假设是如果一个字符串的后继变化数值很低，则可能是一个词的词干，因此，利用后继变化数将词切分为词干和后缀。后继变化数的确定主要通过 4 种方法：①cut off 值方法，通过确定 cut off 值识别边界；②高峰和高地（peak and plateau）法，当后继变化数比前后的值都大时，出现峰值的位置作为切分边界；③完整方法，切出的词必须保持完整性；④熵，应用 cut off 值来定义熵。

——基于文本集的方法。Xu 等[10]提出基于文本集的词干提取方法，假设已知文本中被合并的词形会在文本集中共现。基于该假设，应用词的共现分布测量，而后进行等价类划分。主要使用连接组件算法，对生成的连接组件进行最优划分，最后获得词干见图 7。

Paik 等[10]也提出新的基于共现统计的词干提取方法。其通过检索结果进行算法评估，发现该算法基于统计规律，能够处理不同语种。

Majumder 等[11]提出基于聚类的无监督方法 YASS 算法，定义词对间的字符距离测度方法，通过奖励最长匹配前缀和惩罚错误匹配的算法计算，聚类结果被看作是等价的分类，类的质心作为词干。该算法无需语言知识，能应用于多种语言的词干处理，但是其需要强大的计算能力。

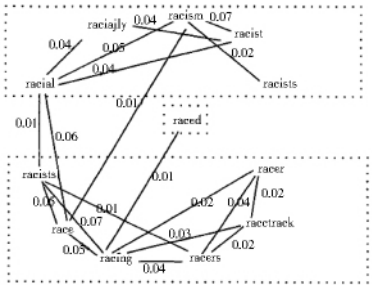

图 7　一个连接组件等价类的最优划分[9]

——基于 n-gram 的方法。Mayfield[12]等提出基于 n-gram 的词干提取方法，主要分析一篇文章中所有的 n-gram 分布。形态始终如一的常量（唯一词干）出现的次数少于变化的部分（前缀和后缀，如“ing”或“able”），采用逆文档词频（IDF）进行计算识别。“jungling”的 4-gram 在 CLEF2002 文本集中的分布如表 1 所示：

表 1　“jungling”的 4-gram 统计分布[12]

n-gram	文档频次	n-gram	文档频次
_jug	681	glin	4 567
jugg	495	ling	55 210
uggl	6 775	ling_	55 210
ggli	3 003		

可见，jugg 的文档频次最低，变化稳定，最可能为词干。该算法通过 8 种欧洲语言进行测试，在处理不同语种的能力上具有一定效果，但它需要消耗大量的内存和存储空间。

——基于 HMM 的方法。Melucci 等[13]提出基于隐马尔科夫模型（HMM）进行词形规范化处理。隐马尔科夫模型是一种用参数表示的用于描述随机过程统计特性的概率模型。随机过程中的每个路径的概率都被计算，以找出最有可能的路径。构成一个词的每个字母都被看作是一种状态。作者将所有可能的状态划分为两组（词根和词尾）和两类初始点（词根）与终点（词根或词尾）。状态间的转换定义了词建立的过程。而对于每个已知词，从起始点到终点有很多可能的路径，其中能够产生切分点（从词根到词尾的转换），切分点前的序列就被作为词干。通过对 5 种欧洲语言进行词干提取和结果评估，发现该算法在处理多种语言上是有效的。

2.1.4　混合的方法　混合的方法主要采用两种或更

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

多的方法综合解决词干提取问题，例如词缀删除和基于词典查找方法的结合。Silva 等[14] 提出基于词典的词干提取算法，实际上是将词缀删除与词典查找结合起来。将词缀删除的规则存储在表中，这个表被存储于最小确定有限自动机中，是词典的有效表示形式。词缀删除算法的实现使用了半自动方法创建的通用规则列表。该方法主要用于处理葡萄牙语，效果较好。

Bonus[15] 提出 Tagalog 算法用于 Tagalog 的词干提取。包括两个阶段：①非取词干阶段，主要进行连字符查询和词典查询，其使用的词典包含 1 575 个词干；②词干提取阶段，通过词典查找删除词缀。该算法先使用规则而后进行词典查找，迭代进行，用于避免过度去词缀的问题。

还有一些方法在现有方法的基础上进行了改进，如 AI-Shammari[16] 将语法知识引入词干提取算法，结合后缀删除方法使用。其发现停用词在识别词干中的作用，将停用词分为两类：有用停用词和无用停用词。算法利用有用停用词识别词性，引入语法知识自动构建三类词典：名词词典、动词词典和形容词词典。而后通过删除停用词，利用最长匹配法进行词缀删除。与 Porter 算法、Lovin 算法相比，它取得了相对较好的效果。Pande 等[17] 考虑了语音因素，认为具有相同词干的词具有近似的发音。首先生成词的 Metaphone 编码，构造出具有相同 Metaphone 编码的词，形成词的列表，而后计算每个词的编辑距离（ED）和包含的最长共有序列（LCS）。基于规则进行词的过滤后，选择具有最大 ED 值和最小 LCS 值的词，进行词干提取。

2.2 词干提取方法的比较分析

基于规则的方法和词典查找法为词干提取方法中的主流技术。它们是该领域中经典的和具有较大应用范围的方法，分别在计算效率和准确性上具有较强优势，研究较为成熟。但是基于规则的方法需要开发者熟悉语言知识，否则无法定义语言形态变化规则。词典查找法受到词典规模、类型、领域的限制。这两种方法具有语种依赖性，对于不规则的、词典外的词形规范也具有局限性。

基于统计的词干提取方法在对词典范围外的未收录词、非规则变形词的处理上具有重要潜力。其不仅无需语言背景知识，而且能够处理多种语言。当前对该类方法的研究较多，但是其具有文本集的领域或训练数据的依赖性，并且对计算能力及存储要求较高，尚处于研究测试阶段。

混合的方法是基于上述方法的局限性而提出的有效方法，融合了多种方法的优势。目前，越来越多的研究者深入挖掘语言的其他特征，如语法、语音等作为词干提取的考虑因素。

目前研究中，并不是每类算法完全独立，具有绝对明显的界限，而是算法在保持自己特色的同时，融合及借鉴部分其他类算法作为补充。例如词典查找法中也结合了少量语言规则；基于规则的方法中将词缀规则存储为简单列表的形式。在基于统计的方法和混合方法中，研究者也考虑了对语言变形规则的提取。

部分算法考虑并解决拼写问题。在英语词干提取的算法中，存在拼写问题，主要包括两方面：一是拉丁文的拼写，另一方面是美式英语和英式英语的区别。处理拼写问题的算法主要采用了词缀或字母转换替代的方法。

综上，四大类方法的优缺点及特点等如表 2 所示：

表 2 词干提取方法比较分析

比较项目	基于规则的方法	词典查找法	基于统计的方法	混合的方法
优点	易于处理常规词形变换，针对性强；计算速度快。	通过词典匹配，结果更为准确；易于扩展和维护；返回结果可能是词典中的有效词。能处理部分不规则变换。	通过统计规律，不受语种限制；无需语言知识，能够处理词典中未包含词和不规则词形变化；能够处理多种语言。	克服单独方法的局限性，融合多种方法的优势，提高词干提取结果的准确率。
缺点	对于不规则变形词具有局限性。	词干识别受词典范围限制，不能识别词典中未包含的词。字典过大时，遍历查找费时，具有词典依赖性。	识别出的词干噪音较大，准确率受选取的文本集影响。	多方法融合，算法流程相对复杂，效率受到一定损失。
特点	通过语言形态中的特定规律去除词缀、保留词干。	利用词形和词干对照词典进行映射，将词的变形转换为词干。	借助词在文本集中的分布规律进行词干提取。	综合两种以上的方法进行词干提取。考虑多因素，如规则、词典、语音、语法等。
开发及维护要求	算法开发规则的建立需要语言形态学背景。	初期需人工或半自动化建立词典或选择权威词典。	受选取的文本集影响，算法消耗较大，需要较强的计算能力和较多的内存或存储空间。	需要多种背景知识支持。

3 主要实现工具

词干提取的经典算法有其主要的实现工具，在形态分析和自然语言处理的相关项目或开发工具当中，词干提取处理被作为重要的功能操作，相关工具提供一种或多种算法。Lovin 算法和 Lancaster 算法有单独实现的工具，包括多种开发语言的版本，Snowball 提供了 Porter 算法及 Lovin 算法。语言形态分析及拼写工具如 Morph Adorner、Hunspell，将词干提取作为其主要实现的功能之一。

3.1 词干提取工具

3.1.1 Lovin 算法实现工具 Lovin 算法的官方网站提供了不同开发语言版本的 Lovin 算法。包括 C 语言、Java 语言和 Perl 语言的版本。

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3.1.2 Lancaster 算法实现工具 算法由 Lancaster 大学的电子计算系于 20 世纪 90 年代开发。最初采用 Pascal 语言开发,随后使用 C 语言和 Java 语言开发,还有 Perl 的版本。Lancaster 算法由一个取词干算法和一组取词干规则构成。此外,还有 Antonio Zamora [18] 提出改进的 Lancaster 算法,扩展了原有算法的词形规则列表,增加了新的规则集,对词干的最小字符数和最大字符数进行了约束等,并提供了 Lancaster 改进算法的 C 语言版本实现工具。

3.1.3 Snowball Snowball [19] 是应用于信息检索的词干提取项目,并实现了一些有用的词干提取功能,支持多种开发语言版本,目前提供 Java、Perl、ANSI(C) Prolog、VB 等多种开发版本。其以 Porter 算法著称,也提供 Lovin 算法实现。Porter 算法实现的程序包很小,代码少于 400 行,运行速度快。该项目还提供多种语言的词干提取算法,包括英语、法语、西班牙语、葡萄牙语、意大利语等。该项目中的词干提取算法被广泛用于现有检索项目和文本处理项目中。

3.2 语言形态分析工具

3.2.1 Hunspell Hunspell [20] 是一个用于拼写检查和形态分析的语言处理工具。它支持 C/C + +、Java、Delphi、Ruby、Perl、Python 等编程语言。Hunspell 提供词干提取的功能,词干提取包含区分大小写和不区分大小写功能。其词干提取功能通过词典实现,使用的词典是由 OpenOffice. org 提供的多语言词典,由词缀和词典文件组成,词典包括 Myspell & Hunspell 词典和 Aspell 词典。OpenOffice、Firefox、Thunderbird 默认的都是使用这个拼写检查工具。Lucene 的 3.5 版最新集成了其词干提取的功能。

3.2.2 MorphAdorner MorphAdorner [21] 基于 Java 语言,主要用于处理一篇文本中的词的形态学修饰。MorphAdorner 提供用于文本修饰的标准拼写、词性标注和原形化的方法。MorphAdorner 也提供文本切分、句子边界识别与名字和地名抽取的功能。两个广泛使用的词干提取算法被用于 MorphAdorner 中,包括 Porter 算法和 Lancaster 算法。MorphAdorner 的词干提取是对原形化功能的简单替代,尝试将一个词减为删除词缀的基础形式。

3.3 集成词干提取功能的其他项目工具

除了词干提取算法的单独实现工具及语言形态分析工具,其作为信息检索和文本处理的必要步骤和重要基础操作,被集成到很多信息检索、主题标引、文本处理的相关项目和工具当中。典型项目包括检索项目 Themis [22]、Lemur [23] 和 Lucene [24],主题标引项目 Kea [25-26]、Maui [27] 以及自然语言处理项目 Stanford CoreNLP [28]、NLTK [29] 等。如表 3 所示:

表 3 集成词干提取功能的其他项目工具

项目类别	项目名称	项目简介	词干提取算法	处理语种	开发语言
信息检索项目	Themis	是一个用于比较自然语言文本的信息检索框架。包括理论检索模型的实现(向量空间模型和基于主题的向量空间模型)。	Porter 算法	英语	Java
	Lemur	是一个用于自然语言模型和信息检索研究的系统。实现基于自然语言模型和向量空间模型的分布式检索:结构化查询、跨语言检索、过滤、聚类等。	Porter 算法、Krovetz 算法	英语、阿拉伯语等	主要采用 C + + 语言开发,支持 C + +、Java、C + + 接口
	Lucene	能够对多语种进行自动标引,搜索引擎 3.4.0 的版本中,包含多语种分析器,如丹麦语、希腊语、德语、英语等。一些分析器中包含轻量级的词干提取功能,如捷克语、匈牙利语等。3.5 版本中包含了 Hunspell 的词干提取功能。	在英语分析器中提供 Krovetz 算法,集成了 Snowball 项目,也能够提供 Snowball 包含的 Porter 算法及其他多语种词干提取算法功能	丹麦语、希腊语、德语、英语、捷克语、匈牙利语等	Java
主题标引项目	Kea	是一个对文本进行关键短语抽取的算法工具,能够提供文本的自由标引或受控标引。	Porter 算法、Lovin 算法	法语、德语和西班牙语	Java
	Maui	是一个多目的自动主题标引开源工具,能够完成以下任务:基于叙词表的关键词分配,主题词标引,利用 Wikipedia 进行主题标引,关键词短语抽取,术语抽取,自动标注等。	Lovin 算法、迭代 Lovin 算法、S - stem 算法、Porter 算法	英语、法语和西班牙语	Java
自然语言处理项目	Stanford-CoreNLP	提供一系列自然语言分析工具,对英文语言文本进行处理,实现词干提取、词性标注、命名实体识别等功能。	主要是删除屈折形态,考虑名词、代词、动词的词形变化。其词干提取利用词典查找和规则相结合的方法实现	英语	Java
	NLTK	是一组自然语言处理的模块工具包,包含句子抽取、标注、解析等功能。其中也包括取词干和原形化功能。	Lancaster 算法、Snowball 的 Porter 算法及多语种算法	英语、葡萄牙语、西班牙语、罗马语、德语、瑞典语、俄语、挪威语、荷兰语、丹麦语、阿拉伯语等	Python

3.4 实现工具的特点分析

在实现工具中,主流词干提取算法是基于规则和基于词典查找原理来实现的,具有较为深入的研究和相对稳定的应用。最为成熟的通用工具还是 Snowball 的 Porter 算法,因为简单、高效,其被单独使用或集成于各种大型检索及文本处理的项目当中,如 Lucence、Solr 等。基于统计或机器学习的方法在现有工具中也有部分应用,作为辅助用于规则的进一步优化和训练,或用于处理未知词的词干提取。

现有工具大多数为用 Java(如 Snowball、Stanford CoreNLP) 、C + +(Lemur、Hunspell) 或 Perl 语言开发的

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

开源工具，具有二次集成和开发的价值。

当前词干提取工具中不仅提供英语的词干提取，还提供多种语言的词干提取功能。如 NLTK、Snowball、Maui 支持葡萄牙语、西班牙语等多语种的词干提取。

4 结 论

词干提取方法及技术发展至今，已较为成熟，但是由于自然语言的复杂性，它仍然存在一些问题。如英语派生形态变化规律性不强，而且变形后的词义和词性具有不可预测性，现有方法还不能很好地处理这部分的变化。其他语种，如西班牙语、葡萄牙语等形态变化更为复杂，对其研究有待深入。虽然有多种不同的词干提取算法，但同一词干提取算法无法处理多种不同语言。

今后，词干提取方法、技术工具的发展趋势主要为：①多语种词干提取。不同语言形态特征不同，针对不同语种的词干提取算法逐步成为研究中的一个难点和热点。②深入语言特征的词干提取。目前，词干提取方法从词的构成发展到结合语法、语音等因素来解决词干提取问题，未来研究将结合更多的语言特征。③增加拼写错误检查，保证词形转换替代的正确性；此外，如果将词干转为有效词将更能提高结果的可用性。④提高准确率和处理速度。现有研究中，研究者提出并实现了多种词干提取方法，综合各方法的长处能达到更好的效果，但同时也要考虑词干处理的效率。因为其往往被用于处理大规模的检索数据，需要保证词干提取的速度和结果的准确性。

参考文献：

[1] Jurafsky D，Martin J H. 自然语言处理综论 [M]. 冯志伟，孙乐，译. 北京：电子工业出版社，2005：38.

[2] The Porter stemming algorithm [OL]. [2011 - 10 - 25]. http://tartarus.org/martin/PorterStemmer/.

[3] Lancaster [OL]. [2011 - 10 - 21]. http://www.comp.lancs.ac.uk/computing/research/stemming/.

[4] Lovin [OL]. [2011 - 10 - 21]. http://www.cs.waikato.ac.nz/~eibe/stemmers/

[5] Dawson J L. Suffix removal for word conflation [J]. Bulletin of the Association for Literary & Linguistic Computing，1974，2(3)：33 - 46.

[6] Giridhar N S，Prema K V，Reddy N V S. A prospective study of stemming algorithms for Web text mining [J]. Ganpat University Journal of Engineering & Technology，2011，11(1)：28 - 34.

[7] Krovetz R. Viewing morphology as an inference process [J]. Artificial Intelligence，2000，118(1)：227 - 294.

[8] Hafer M A，Weiss S F. Word segmentation by letter successor varieties [J]. Information Storage and Retrieval，1974，10(11/12)：371 - 385.

[9] Xu Jinxi，Croft W B. Corpus-based stemming using co-occurrence of word [J]. ACM Transactions on Information Systems Journal. 1998，16(1)：61 - 81.

[10] Paik J H，Pal D，Parui S K. A novel corpus-based stemming algorithm using co-occurrence statistics [C]//Ma Weiying，Nie Jianyun，Baeza-Yates R A. Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York：ACM，2011：863 - 872.

[11] Majumder P，Mitra M，Parui S K，et al. YASS：Yet another suffix stripper [J]. ACM Transactions on Information Systems，2007，25(4)：18 - 38.

[12] Mayfield J，McNamee P. Single n-gram stemming [C]//Efthimiadis E N，Dumais S T，Hawking D，et al. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. New York：ACM，2003：415 - 416.

[13] Melucci M，Orio N. A novel method for stemmer generation based on hidden Markov models [C]//Kraft D，Frieder O，Hammer J，et al. Proceedings of the Twelfth International Conference on Information and Knowledge Management. New York：ACM，2003：131 - 138.

[14] Silva G，Oliveira C. A lexicon-based stemming procedure [C]//Mamede N J，Trancoso I，Baptista J，et al. Proceedings of the 6th International Conference on Computational Processing of the Portuguese Language. Berlin：Springer-Verlag，2003：159 - 166.

[15] Bonus D E J [OL]. [2011 - 10 - 22]. http://www.dlsu.edu.ph/research/centers/adric/nlp/downloads/1NNLPRS/Paper%5Cbonus.pdf.

[16] Al-Shammari E T. Towards an error-free stemming [OL]. [2011 - 10 - 21]. http://www.iadis.net/dl/final_uploads/200812R027.pdf.

[17] Pande B P，Dhami H S. Application of natural language processing tools in stemming [J]. International Journal of Computer Application，2011，27(6)：14 - 19.

[18] Modifications to the Lancaster stemming algorithm [OL]. [2011 - 10 - 21]. http://www.scientificpsychic.com/paice/paice.html

[19] Snowball [OL]. [2011 - 10 - 21]. http://snowball.tartarus.org/.

[20] Hunspell [OL]. [2011 - 10 - 25]. http://hunspell.sourceforge.net/.

[21] MorphAdoner [OL]. [2011 - 10 - 21]. http://morphadorner.northwestern.edu/morphadorner/.

[22] Themis [OL]. [2011 - 10 - 19]. http://code.google.com/p/ir-themis/.

[23] Lemur [OL]. [2011 - 10 - 19]. http://www.lemurproject.org/.

（下转第 142 页）

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

慧最终仍都变成了期刊等的财产，成为“信息封建贵族”出版商的摇钱树，纳税人根本无法享受到其纳税支持的科学研究所产生的成果。

随着信息封建主义的日益扩张，学者们渴望信息的免费流动及其畅通获取的文化氛围受到破坏，终于促使学术界的觉醒，他们以联合起来抵制 RWA 等实际行动对出版商提出抗议。二者矛盾的升级让出版商们无懈可击的地位在逐渐动摇，毕竟商业出版商需要学者远胜于学者需要出版商。两家最有声望的科学期刊出版商 Science 和 Nature，也公开表示愿意对美国公共获取政策伸出橄榄枝。

同在美国，一方面是 RWA 遭撤回，另一方面则是《联邦研究成果公共获取法案》(*Federal Research Public Access Act*，FRPAA) 的再次提交[8]。FRPAA 作为迄今为止美国提出的最强有力的开放获取提案，于 2012 年再次被提交说明：向公共资助信息的开放性、透明性和可获取性的挺进势不可挡。

英国作为学术论文的主要输出国，科学与大学部长 D. Willetts 明确表明政府的态度是：公众出资赞助的研究，其成果应该由公众共享，而且这个原则不局限于学术领域。Willetts 同时关注到美国的发展，甚至已经准备好要在欧盟相关会议上提起开放获取议题[9]。

2011 年 11 月召开的柏林第 9 次开放获取会议关注如何将人文社科和自然科学领域的研究成果，快速、免费地传递到学者、学生、创新者和公众手中[10]。

实现研究成果的开放获取，学者和研究人员才能在及时、无障碍、在线访问等最有效的服务中，快速获取他人的研究成果，加快研究发现的速度，培养研究创新能力，最终无限扩大科研预想的效果，提升科学与学术研究共享动能。

参考文献：

[1] The cost of knowledge [EB/OL]. [2012 - 04 - 25]. http://the-costofknowledge.com/.

[2] Williams C. District dispatch: Trying to roll back the clock on open access: Research Works Act introduced [EB/OL]. [2012 - 03 - 30]. http://connect.ala.org/node/165002.

[3] Taylor M. Academic publishers have become the enemies of science: The US Research Works Act would allow publishers to line their pockets by locking publicly funded research behind paywalls [EB/OL]. [2012 - 04 - 15]. http://www.guardian.co.uk/science/2012/jan/16/academic-Publishers-enemies-science.

[4] Gowers W T. 爱思唯尔的衰落——我在其中的角色 [J]. 张智民，译. 数学文化，2012，3(1)：91 - 92.

[5] The department for business，innovation and skills，innovation and research strategy for growth [EB/OL]. [2012 - 04 - 10]. http://www.bis.gov.uk/assets/biscore/innovation/docs/i/11-1387-innovation-and-research-strategy-for-growth.pdf.

[6] 达沃豪斯，布雷斯韦特. 信息封建主义 [M]. 刘雪涛，译. 北京：知识产权出版社，2005.

[7] 冯晓青. 知识产权法的公共领域理论 [J]. 知识产权，2007(3)：3 - 11.

[8] Call to action: Tell congress you support the bipartisan Federal Research Public Access Act [EB/OL]. [2012 - 05 - 28]. http://www.taxpayeraccess.org/issues/frpaa/frpaa_action/FRPAA2012.shtml.

[9] Stephen. Willetts' speech on open access: Analysis [EB/OL]. [2012 - 05 - 28]. http://occamstypewriter.org/scurry/2012/05/03/willetts-speech-on-open-access-analysis/.

[10] Joseph H. The impact of open access on research and scholarship: Reflections on the Berlin 9 open access conference [EB/OL]. [2012 - 05 - 28]. http://crln.acrl.org/content/73/2/83.full.

〔作者简介〕易晓阳，女，1979 年生，副研究馆员，发表论文 13 篇，参编著作 2 部。

(上接第 115 页)

[24] Lucene [OL]. [2011 - 10 - 19]. http://lucene.apache.org/java/docs/index.html.

[25] Kea [OL]. [2011 - 10 - 19]. http://www.nzdl.org/Kea/.

[26] Witten I H，Paynter G W，Frank E，et al. KEA: Practical Automatic keyphrase extraction [C] //Rowe N，Fox E A. Proceedings of the Fourth ACM Conference on Digital Libraries. New York: ACM，1999: 254 - 255.

[27] Maui [OL]. [2011 - 10 - 19]. http://code.google.com/p/maui-indexer/.

[28] Stanford CoreNLP [OL]. [2011 - 10 - 21]. http://nlp.stanford.edu/software/corenlp.shtml.

[29] NLTK [OL]. [2011 - 10 - 21]. http://www.nltk.org/.

〔作者简介〕吴思竹，女，1981 年生，助理研究员，博士，发表论文 8 篇；钱　庆，男，1970 年生，研究员，所长助理，硕士，发表论文 20 余篇；胡铁军，男，1949 年生，研究员，硕士生导师，发表论文 20 余篇；李丹亚，女，1954 年生，研究员，硕士生导师，发表论文 20 余篇；李军莲，女，1972 年生，副研究馆员，硕士，发表论文 10 余篇；洪　娜，女，1980 年生，助理研究员，博士，发表论文 10 余篇。

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net