北京大学学报(自然科学版) 第 52 卷 第 1 期 2016 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 52, No. 1 (Jan. 2016)
doi: 10.13209/j.0479-8023.2016.024

基于互信息改进算法的新词发现对中文分词系统改进

杜丽萍 李晓戈† 于根 刘春丽 刘睿

西安邮电大学, 西安 710121; †通信作者, E-mail: lixg@xupt.edu.cn

摘要 提出一种非监督的新词识别方法。该方法利用互信息(PMI)的改进算法——PMI^k 算法与少量基本规则相结合, 从大规模语料中自动识别 2~n 元网络新词(n 为发现的新词最大长度, 可以根据需要指定)。基于 257 MB 的百度贴吧语料实验, 当 PMI^k 方法的参数为10 时, 结果精度达到 97.39%, 比 PMI 方法提高 28.79%, 实验结果表明, 该新词发现方法能够有效地从大规模网络语料中发现新词。将新词发现结果编纂成用户词典, 加载到汉语词法分析系统 ICTCLAS 中, 基于10 KB 的百度贴吧语料实验, 比加载用户词典前的分词结果准确率、召回率和 F 值分别提高 7.93%, 3.73%和 5.91%。实验表明, 通过进行新词发现能有效改善分词系统对网络文本的处理效果。

关键词 新词识别; 未登录词; 互信息; PMI 改进算法; 中文分词

中图分类号 TP391

New Word Detection Based on an Improved PMI Algorithm for Enhancing Segmentation System

DU Liping, LI Xiaoge†, YU Gen, LIU Chunli, LIU Rui

School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121;
† Corresponding author, E-mail: lixg@xupt.edu.cn

Abstract This paper presents an unsupervised method to identify internet new words from the large scale web corpus, which combines with an improved Point-wise Mutual Information (PMI), PMI^k algorithm, and some basic rules. This method can recognize internet new words with length from 2 to n (n is any number as needed). Experimented based on 257 MB Baidu Tieba corpus, the precision of proposed system achieves 97.39% when the parameter value of PMI^k algorithm is equal to 10, and the precision increases 28.79%, compared to PMI method. The results show that proposed system is significant and efficient for detecting new word from the large scale web corpus. Compiling the results of new word discovery into user dictionary and then loading the user dictionary into ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System), experimented with 10 KB Baidu Tieba corpus, the precision, the recall and F-measure were promoted 7.93%, 3.73% and 5.91% respectively, compared with ICTCLAS. The result show that new word discovery could improve the performance of segmentation for web corpus significantly.

Key words new word recognition; unknown word; PMI; improved PMI algorithm; Chinese word segmentation

随着信息时代的发展与科学技术的进步, 大量网络新词不断涌现, 使得分词结果中存在大量的“散串”, 严重影响分词系统处理网络文本的效果, 新词识别已经成为提高分词效果的瓶颈[1]。

对于网络上出现的新词汇, 例如近日在网上热传的“APEC 蓝”、“Duang”、“一带一路”、“单肾贵

国家自然科学基金(61373116)、陕西省普通高等学校重点学科专项资金(112-1602)和西安邮电大学研究生创新基金(ZL2013-31)资助
收稿日期: 2015–06–07; 修回日期: 2015–09–14; 网络出版日期: 2015–09–29

族”和“花样作死”等词语，一般的识别方法是基于大规模语料库，由机器根据某个统计量自动抽取出候选新词，再由人工筛选出正确的新词[2]。Pecina 等[3]采用 55 种不同的统计量进行 2 元词汇识别实验，结果表明，PMI 算法是最好的衡量词汇相关度的算法之一。通常情况下，PMI 方法能够很好地反映字串之间的结合强度，但缺点是过高地估计低频且总是相邻出现的字串间的结合强度[3–4]。例如，“啰”和“嗦”、“蝙”和“蝠”等在语料库中低频且总是相邻出现，这些字串的 PMI 值非常高，包含这些低频字串的垃圾串的 PMI 值也非常高，例如“很啰”和“嗦”、“的蝙”和“蝠”等。针对此问题，研究者将 PMI 方法与其他方法相结合进行新词发现研究。文献[5–7]均采用 PMI 方法与 log-likelyhood 方法相结合进行新词识别。梁颖红等[8]利用 PMI 方法衡量字串间的结合强度，结合 NC-value 方法融入词语上下文信息来提高 3 个字以上长新词的抽取精度。何婷婷等[9]采用互信息方法 F-MI 抽取结构简单的质词。孙继鹏等[10]提出一种语言文法信息与互信息相结合的新词识别方法。Pazienza 等[11]提出使用 PMI^2 和 PMI^3 的方法改进 PMI 方法来识别新词。Bouma[12]通过向 PMI 方法中引进 k 个联合概率因子，改善 PMI 方法的缺点，这种改进的 PMI 方法称为 PMI^k 方法。杜丽萍等[13]通过抽象语料库中低频且总是相邻出现字串的数学特征，从理论上证明，当向 PMI 方法中引进 3 个及以上的联合概率因子时，PMI^k 方法能够克服 PMI 方法的缺点。

目前，常用的分词方法主要有 3 种：基于词表的分词方法、基于统计模型的分词方法和基于统计方法与规则方法相结合的分词方法[2]。3 种方法均有优点，但也存在不足：基于词表的分词方法效率高，但对新词的识别能力不足[14]；基于规则的方法很难涵盖所有的语言现象[2]，尤其对网络语料的处理能力非常有限；基于统计模型的分词方法重点在于解决自动分词的歧义分词问题，但需要人工标注训练语料，且受训练语料领域的限制。ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System)是基于隐马尔科夫统计模型(HMM, Hidden MarKov Model)进行分词的广受好评的中文分词系统，ICTCLAS2002 版在国内 973 评测中综合第一名，经过 15 年打造，ICTCLAS2015 版又增加了新词自动识别功能。

本文在杜丽萍等[13]的定理 1 和定理 2 基础上，采用非监督的基于 PMI^k 与少量的基本规则相结合的方法，从大规模网络语料中自动识别新词，并对 ICTCLAS2002 版分词系统进行改进，对比改进后的 ICTCLAS2002 分词系统与 ICTCLAS2002 和 ICTCLAS2015 版的分词效果。

1 分词系统改进

1.1 改进分词系统框架

分词系统改进主要分为两个阶段：1)基于大规模语料库进行新词发现；2)用新词发现结果编纂用户词典，加载到分词系统中。图 1 为改进的分词系统的流程。

1.2 基于 PMI 改进方法的新词发现

定义 1 PMI^k 算法[12]定义如下：

$$\mathrm{PMI}^k(x, y) = \log \frac{p^k(x, y)}{p(x)p(y)},\ k \in N^+,$$

其中，$p(x)$和 $p(y)$分别表示字串 x 和 y 的概率，$p(x, y)$表示字串 x 和 y 的联合概率，$\mathrm{PMI}^k(x, y)$表示字串 x 和 y 的相关度，也称 PMI^k 值。特殊地，当 k=1 时，PMI^k 方法即 PMI 方法。

新词发现过程主要分为 4 个阶段：1) 确定 2 元待扩展种子；2) 将 2 元待扩展种子扩展至 2~n 元；3) 过滤候选新词；4)人工判定。算法的步骤如下。

步骤 1 从 4 元字串中确定出 2 元的待扩展种子。对于每一个 4 元字串 $w_{i-1}w_iw_{i+1}w_{i+2}$，计算中间两元字串 w_iw_{i+1} 和前两元字串 $w_{i-1}w_i$ 的 PMI^k 值之和的平均值 mean_1 以及中间两元字串 w_iw_{i+1} 和后两元字串 $w_{i+1}w_{i+2}$ 的 PMI^k 值之和的平均值 mean_2。计算公式如下：

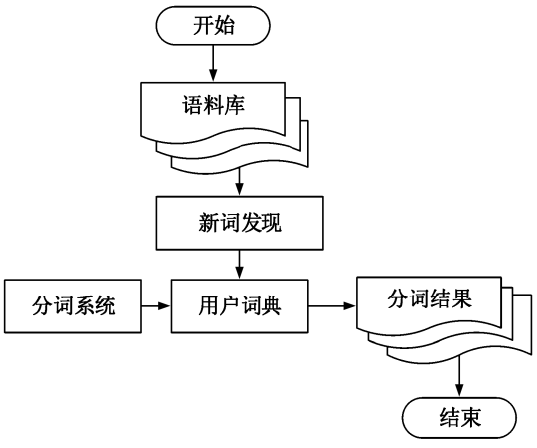

图 1 改进分词系统流程

Fig. 1 Flow chart of the improved segmentation system

$$\text{mean}_1=\frac{1}{2}(\text{PMI}^k(w_i,w_{i+1})+\text{PMI}^k(w_{i-1},w_i))，$$

$$\text{mean}_2=\frac{1}{2}(\text{PMI}^k(w_i,w_{i+1})+\text{PMI}^k(w_{i+1},w_{i+2}))。$$

对于 4 元字串 $w_{i-1}w_iw_{i+1}w_{i+2}$，如果满足

$$\text{PMI}^k(w_i,w_{i+1})>\text{PMI}^k(w_{i-1},w_i)+\text{mean}_1，$$

$$\text{PMI}^k(w_i,w_{i+1})>\text{PMI}^k(w_{i+1},w_{i+2})+\text{mean}_2，$$

则认为字串 w_iw_{i+1} 是一个词或者词的一部分的概率较大，即 2 元字串 w_iw_{i+1} 为待扩展种子，执行步骤 2；否则，认为字串 w_i 和 w_{i+1} 各自成词或是词的边界的概率较大，字串 w_iw_{i+1} 的串频减 1。

步骤 2　将 t 元字串扩展至 $t+1$ 元字串，其中 $t\in[2,n-1]$。取出待扩展字串 $w_i,...,w_{i+t-1}$ 的前一元 w_{i-1} 和后一元 w_{i+t}，分别计算 $\text{PMI}^k(w_{i-1},w_i,...,w_{i+t-1})$ 和 $\text{PMI}^k(w_i,...,w_{i+t-1},w_{i+t})$。有如下两种可能性。

1)如果 $\text{PMI}^k(w_{i-1}, w_i, \ldots, w_{i+t-1})>\text{PMI}^k(w_i, \ldots, w_{i+t-1})$，则认为把字串 $w_i,...,w_{i+t-1}$ 扩展成 $w_{i-1},...,w_{i+t-1}$ 的概率大于扩展成 $w_i,...,w_{i+t}$ 的概率，故向前扩展。计算 $\text{mean}=\frac{1}{2}(\text{PMI}^k(w_{i-1}, w_i,..., w_{i+t-1})+\text{PMI}^k(w_i, \ldots, w_o, w_{o+1}, ..., w_{i+t-1}))$，其中 $o=i$ 或 $o=i+t-2$。如果满足

$$\text{PMI}^k(w_{i-1},w_i,...,w_{i+t-1})+\text{mean}\geqslant \text{PMI}^k(w_i,...,w_o,w_{o+1},...,w_{i+t-1})，$$

则把 t 元字串 $w_i,...,w_{i+t-1}$ 扩展成 $t+1$ 元字串 $w_{i-1},...,w_{i+t-1}$，$t=t+1$，依次迭代，执行步骤 2；否则，输出 t 元字串 $w_i,...,w_{i+t-1}$，执行步骤 3。

2)如果 $\text{PMI}^k(w_{i-1},w_i,...,w_{i+t-1})\leqslant\text{PMI}^k(w_i,\ldots,w_{i+t-1},w_{i+t})$，则认为把字串 $w_i,...,w_{i+t-1}$ 扩展成 $w_i,...,w_{i+t}$ 的概率大于扩展成 $w_{i-1},\ldots,w_{i+t-1}$ 的概率，故向后扩展。计算 $\text{mean}=\frac{1}{2}(\text{PMI}^k(w_i,...,w_o,w_{o+1},...,w_{i+t-1})+\text{PMI}^k(w_i, \ldots, w_{i+t-1}, w_{i+t}))$，其中 $o=i$ 或 $o=i+t-2$。如果满足

$$\text{PMI}^k(w_i,...,w_{i+t-1},w_{i+t})+\text{mean}\geqslant \text{PMI}^k(w_i,...,w_o,w_{o+1},...,w_{i+t-1})，$$

则把 t 元字串 $w_i,...,w_{i+t-1}$ 扩展成 $t+1$ 元字串 $w_i,...,w_{i+t}$，$t=t+1$，依次迭代，执行步骤 2；否则，输出 t 元字串 $w_i,...,w_{i+t-1}$，执行步骤 3。

步骤 3　利用可存在性过滤规则。如果 t 元字串 $w_i,...,w_{i+t-1}$ 的串频小于阈值 T，则退出算法；否则，执行步骤 4。

步骤 4　利用停用词过滤规则。如果 t 元字串 $w_i,...,w_{i+t-1}$ 的任意一个子串包含在停用词集合中，则退出算法；否则，按 $\text{PMI}^k(w_i,...,w_o,w_{o+1},...,w_{i+t-1})$ 值降序地把字串 $w_i,...w_{i+t-1}$ 加入候选新词链 L，执行步骤 5。

步骤 5　根据核心词表，过滤候选新词链 L 上的核心词汇，执行步骤 6。

步骤 6　人工判定。

2　实验及结果分析

2.1　实验数据

1) 257 MB (约 1000 万字)百度贴吧语料，用于网络新词发现。

2) 停用词典：包含 702 个停用词(选自哈尔滨工业大学停用词表)，用于过滤候选新词结果中的垃圾串。

3) ICTCLAS 核心词典：共收集 79836 个词语，是目前比较规范的词典之一，用于过滤候选新词结果中的核心词汇，以便得到新词。

4) 10 KB 百度贴吧语料，用于测试分词系统改进的效果。

2.2　新词实验及结果

黄昌宁等[15]指出，99%以上的词长都在五字及五字以下，故本实验设定抽取的最大词长 n 等于 5。

由于难以统计 257 MB 百度贴吧语料中的全部新词，所以只采用准确率作为衡量新词发现方法的评测标准。准确率计算公式为

$$\text{准确率}=\frac{\text{正确新词条数}}{\text{新词条数}}\times100\%。$$

在 PMI^k 方法的参数 k 取 1~10 之间 10 个正整数值时，分别进行实验，图 2 描述随着 k 值变化的准确率变化趋势。

表 1 列举 PMI^k 方法的参数 k 取 1~10 之间 10 个正整数值时，新词结果的前 20 条。

2.3　改进分词系统实验及结果

实验设计如下。实验一：基于 ICTCLAS2002 版分词系统进行实验；实验二：基于 ICTCLAS2015 版分词系统进行实验；实验三：加载用户词典到 ICTCLAS2002 版分词系统中进行实验。采用准确率、召回率和 F 值 3 个指标来衡量分词系统的性能，计算公式如下：

$$\text{准确率}=\frac{\text{切分正确的词数目}}{\text{切分出的总词数}}\times100\%，$$

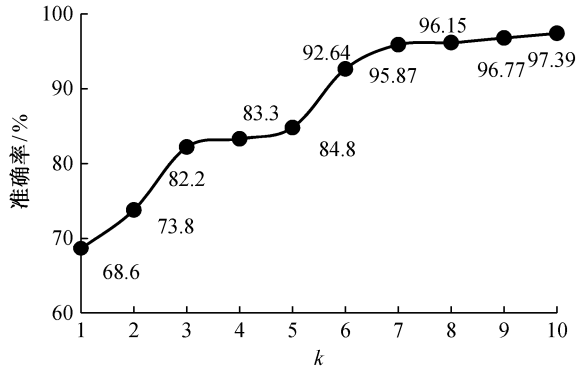

图 2 随着 k 值变化的准确率变化趋势

Fig. 2 Precision trends with the value of k changes

$$召回率=\frac{切分正确的词数目}{文本包含的总词数}\times 100\%,$$

$$F=\frac{准确率\cdot 召回率\cdot 2}{准确率+召回率}\times 100\%。$$

针对 10 KB 百度贴吧测试语料进行上述实验，实验结果如表 2 所示，“切分出总词数”表示分词系统切分出的字串总数目，“识别新词数目”表示分词结果中包含的正确的新词数目。

表 3 列举 10 KB 百度贴吧语料中 3 个例句分别在实验一、实验二和实验三中的结果。

例 1 让我这个菜鸟都有点情何以堪啊!

例 2 这个镜头在变形金刚刚出来时候不是就被喷了么?

例 3 小正太，你好。

2.4 结果分析

从图 2 可以看出，准确率随 k 值增大而增大且逐渐趋于 100%。$k=3$ 时的准确率比 $k=1$ 时提高 13.6%，$k=10$ 时的准确率比 $k=1$ 时提高28.79%。因此，当 PMI^k 方法的参数 $k\geqslant 3$ 时，PMI^k 方法能明显改善新词识别的效果。

由表 1 看出，当 PMI^k 方法的参数 $k\geqslant 3$ 时，新词识别结果与 $k=1$ 和 $k=2$ 时差异较大。在 $k=1$ 和 $k=2$ 的结果中，排名在前的字串中均包含低频的字或词，例如垃圾串“晦涩难”、“非贪婪”、“徽太尉”、“吧头衔”中分别包含“晦涩”、“婪”、“徽”、“衔”等低频字串，且这些字串的搭配词语固定。该现象反映出 PMI 方法和 PMI^2 方法对低频共现字串敏感的缺点。在 $k\geqslant 3$ 的结果中，均没有出现低频共现字串，说明 $k\geqslant 3$ 时 PMI^k 方法克服了 PMI 方法的缺点，PMI^k 方法能有效识别新词。

从表 2 可以看出，相对 ICTCLAS2002 加载用户词典前，ICTCLAS2002 加载用户词典后分词系统识别出的新词数目增加 149 个，准确率、召回率和 F 值也分别提高 7.93%，3.37%和 5.91%。结果表明，增加用户词典后，ICTCLAS2002 分词系统处理网络语料的效果有明显改善。相对 ICTCLAS2015 分词系统，ICTCLAS2002 加载用户词典后分词系统识别出的新词数目增加了 124 个，准确率、召回率和 F 值也分别提高 6.7%，3.1% 和 4.96%。

表 3 中，针对例 1，ICTCLAS 2002 和 ICTCLAS2015

表 1 前 20 条实验结果

Table 1 First 20 experimental results

k	实验结果
1	晦涩难, 非贪婪, 周子琦, 嘤嘤, 金针菇, 啰嗦, 耦合度, 肝肠, 蜀黍, 吧头衔, 矢量图, 抠脚大, 瞅瞅, 衲法号, 可理喻, 天答辩, 烫烫, 鼎鼎, 仔细观察, 彬彬
2	南海保镖, 赫卡特, 青年范儿, 刘易雯, 徽太尉, 满智勇, 寒云似雾, 童鞋, 叼叼, 云似雾, 冒险岛, 迭代器, 吐槽, 蜀黍, 楠绾绾, 锐英源, 蛋疼, 莱克斯, 御坂, 肝肠
3	真朱, 寒云, 大神, 蛋疼, 窗体, 良化, 百度, 楼主, 控件, 菜鸟, 童鞋, 吐槽, 渡娘, 膜拜, 递归, 炮姐, 余贺, 坑爹, 尼玛, 傲娇
4	真朱, 大神, 楼主, 窗体, 百度, 良化, 控件, 寒云, 蛋疼, 菜鸟, 贴吧, 渡娘, 童鞋, 源码, 帖子, 点击, 递归, 链接, 吐槽, 线程
5	大神, 楼主, 真朱, 窗体, 控件, 百度, 良化, 寒云, 蛋疼, 菜鸟, 贴吧, 源码, 点击, 帖子, 线程, 渡娘, 链接, 童鞋, 微软, 递归
6	大神, 楼主, 真朱, 控件, 窗体, 百度, 良化, 寒云, 蛋疼, 菜鸟, 贴吧, 源码, 线程, 点击, 帖子, 链接, 渡娘, 次元, 微软, 神马
7	大神, 楼主, 真朱, 控件, 窗体, 百度, 良化, 寒云, 蛋疼, 贴吧, 菜鸟, 源码, 线程, 点击, 帖子, 链接, 次元, 渡娘, 微软, 神马
8	大神, 楼主, 真朱, 控件, 窗体, 百度, 良化, 寒云, 蛋疼, 贴吧, 源码, 菜鸟, 线程, 点击, 帖子, 链接, 次元, 微软, 神马, 报错
9	大神, 楼主, 真朱, 控件, 窗体, 百度, 良化, 寒云, 贴吧, 蛋疼, 源码, 菜鸟, 线程, 点击, 帖子, 链接, 次元, 神马, 报错, 微软
10	大神, 楼主, 真朱, 控件, 窗体, 百度, 良化, 寒云, 贴吧, 蛋疼, 源码, 线程, 菜鸟, 点击, 帖子, 链接, 次元, 报错, 神马, 微软

表 2　实验结果
Table 2　Experimental results

实验编号	切分出总词数	识别新词数目	准确率/%	召回率/%	*F*/%
实验一	4260	1	89.81	95.65	92.64
实验二	4230	26	91.04	96.28	93.59
实验三	4067	150	97.74	99.38	98.55

表 3　实验结果举例
Table 3　Example of experimental results

实验编号	实验结果
ICTCLAS2002	让/ 我/ 这个/ 菜/ 鸟/ 都/ 有/ 点/ 情/ 何以/ 堪/ 啊! 这个/ 镜头/ 在/ 变形/ 金刚/ 刚/ 出来/ 时候/ 不/ 是/ 就/ 被/ 喷/ 了/ 么/ ？ 小/ 正/ 太/ , / 你/ 好/ 。
ICTCLAS2015	让/ 我/ 这个/ 菜/ 鸟/ 都/ 有/ 点/ 情/ 何以/ 堪/ 啊! 这个/ 镜头/ 在/ 变形/ 金/ 刚刚/ 出来/ 时候/ 不/ 是/　就/ 被/ 喷/ 了/ 么/ ？ 小正太/ , / 你/ 好/ 。
ICTCLAS2002 加载用户词典	让/ 我/ 这个/ 菜鸟/ 都/ 有/ 点/ 情/ 何以/ 堪/ 啊! 这个/ 镜头/ 在/ 变形金刚/ 刚/ 出来/ 时候/ 不/ 是/ 就/ 被/ 喷/ 了/ 么/ ？ 小正太/ , / 你/ 好/ 。

分词系统均把新词“菜鸟”切分为“菜/ 鸟”; ICTCLAS2002 加载用户词典(词典中包含新词“菜鸟”)后，分词系统把新词“菜鸟”切分为一个词。针对例 2，ICTCLAS2002 分词系统把新词“变形金刚”切分为“变形/ 金刚”; ICTCLAS2015 分词系统分词把“变形金”切分为一个词，把“变形金刚”中的“刚”和它后面的“刚”结合起来切分为“刚刚”; ICTCLAS2002 加载用户词典(词典中包含新词“变形金刚”)后，分词系统把新词“变形金刚”切分为一个词。针对例 3，ICTCLAS2002 分词系统把新词“小正太”切分为“小/ 正/ 太”; ICTCLAS2015 和 ICTCLAS2002 加载用户词典(词典中包含新词“小正太”)后分词系统把新词“小正太”切分为一个词。从 10 KB 百度贴吧测试语料的分词结果来看，主要有 3 种情况: 1) ICTCLAS2002 和 ICTCLAS2015 分词系统在遇到新词时，大多情况下均是将新词切分为多个“散串”，如例1, ICTCLAS2002 加载包含这些新词的用户词典之后，这些新词均能被正确切分; 2) ICTCLAS2015 分词系统自动识别出新词不正确，导致句子中其他词的分词结果不正确，如例 2 中把“变形金”当做一个词，导致“变形金刚”后面的“刚”和“变形金刚”中的“刚”结合起来切分为“刚刚”; 3) 在 ICTCLAS2002 把新词切分为多个“散串”时，ICTCLAS2015 和 ICTCLAS2002 加载用户词典后的分词系统正确切分出新词，如例3。结果表明，通过加载用户词典改进分词系统是一种可靠有效的方法。

3　结语

本文基于 257 MB 百度贴吧语料，验证了 PMI^k 方法的参数 k 取值大于等于 3 时，能够克服 PMI 方法的缺点，并通过调整新词发现算法中的参数来提高长度大于 2 元的新词识别率。最后，验证了基于加载用户词典来改进分词系统是有效可行的方法。下一步工作是研究 PMI^k 方法的参数 k 取值与语料库规模、语料特征等因素的关系，找出一种自适应地确定参数 k 值的方法，提高新词识别效果，进一步增强分词系统处理 Web 文本的能力。

参考文献

[1] 张海军, 史树敏, 朱朝勇, 等. 中文新词识别技术综述. 计算机科学, 2010, 37(3): 6–12

[2] 宗成庆. 统计自然语言处理. 北京: 清华大学出版社, 2008: 103–146

[3] Pecina P, Schlesinger P. Combining association measures for collocation extraction // Proceeding Soft of the 21th International Conference on Computational Linguisticsand 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL2006). Sydney, 2006: 651–658

[4] 刘华. 一种快速获取领域新词语的新方法. 中文信息学报, 2006, 20(5): 17–23

[5] 刘建舟, 何婷婷, 骆昌日. 基于语料库和网络的新词自动识别. 计算机应用, 2004, 24 (7): 132–134

[6] 韩艳, 林煜熙, 姚建明. 基于统计信息的未登录词的扩展识别方法. 中文信息学报, 2009, 23(3): 24–30

[7] Patrick P, Lin D K. A statistical corpus-based term extractor // Stroulia E, Matwin S. lecture notes in artificial intelligence. London, 2001: 36–46

[8] 梁颖红, 张文静, 周德福. 基于混合策略的高精度长术语自动抽取. 中文信息学报, 2009, 23(6): 26–30

[9] 何婷婷, 张勇. 基于质子串分解的中文术语自动抽取. 计算机工程, 2006, 32(23): 188–190

[10] 孙继鹏, 贾民, 刘增宝. 一种面向文本的概念抽取方法研究. 计算机应用与软件, 2009, 26(9): 28–30

[11] Pazienza M T, Pennnacchiotti M, Zanzotto F M. Terminology extraction: an analysis of linguistic and statistical approaches. Berlin: Springer-Verlag, 2005: 255–279

[12] Bouma G. Normalized (pointwise) mutual information in collocation extraction // Proc Boennial GSCL Conference 2009, Meaning: Processing Texts Automatically. Tubingen, 2009: 31–40

[13] 杜丽萍, 李晓戈, 周元哲, 等. 互信息改进方法在术语抽取中的应用. 计算机应用, 2015, 35(4): 996–1000, 1005

[14] 莫建文, 郑阳, 首照宇, 等. 改进的基于词典的中文分词方法. 计算机工程与设计, 2013, 34(5): 1802–1807

[15] 黄昌宁, 赵海. 中文分词十年回顾. 中文信息学报, 2007, 21(3): 8–19