

· 博士论文 ·

文章编号: 1000—3428(2004)20—0019—03

文献标识码: A

中图分类号: TP311

基于概率统计技术和规则方法的新词发现

贾自艳^{1,2}, 史忠植¹

(1. 中国科学院计算技术研究所, 北京 100080; 2. 中国科学院研究生院, 北京 100039)

摘 要: 新词/短语的识别是自然语言处理、信息检索和机器翻译等领域的一项基础研究。该文分析了已有短语抽取技术, 并结合汉语特点, 提出了基于概率统计技术和规则方法相结合的概念抽取方法。该方法包括高效的“二元语法”统计模型、统计算法、统计选词策略、丰富的规则知识和规则过滤算法。实验证明该方法适用于从大规模语料库中自动高效地发现新词/短语。

关键词: 新词发现; 短语抽取; 二元语法; 语料库

Probabilistic Techniques and Rule Methods for New Word Discovery

JIA Ziyang^{1,2}, SHI Zhongzhi¹

(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080;

2. Graduate School of Chinese Academy of Sciences, Beijing 100039)

[Abstract] New word and phrase discovery is basic research in fields of NLP, IR and MT. The paper gives the method based on probabilistic techniques and rules for new word discovery via analyzing the current techniques of phrase extraction and combining the specialties of Chinese. This method includes the “bi-gram” probabilistic model, the statistical algorithm, the rich rules and rule-based algorithm for word filtering. Experiments show that this technique is fit for automatically and effectually extracting new words/phrases from large corpora.

[Key words] New word discovery; Phrase extraction; Bi-gram; Corpus

1 概述

新词/短语的识别是自然语言处理、信息检索、文本挖掘和机器翻译等领域的一项基础研究。从特定领域的语料库中发现的新词语、新概念可以丰富人类语言知识(词典), 帮助解决一些歧义切分的问题, 提高汉语分词的准确度; 另外新词语常常表达更为精确的概念, 这样可以提高以词为特征项的文本向量空间模型(VSM)的表达能力。文献[1]通过二元语法抽取词组短语对文本向量进行降维, 以克服计算的复杂性。文献[2]使用二元语法来提高文本特征向量的质量, 进而提高了分类的性能。

“基于规则方法和基于语料库方法是计算语言学论著中经常提到的两个术语。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。另外, 自然语言远不是一个精心规划的系统, 我们难以用一套规则去准确地预测真实文本中所出现的各种变异, 因此应当用基于语料库的统计方法来研究自然语言。二者都各有优缺点, 基于统计的方法不受领域限制, 速度很快, 容易实现, 符合当前自然语言处理面向大规模实用语料的发展趋势, 但质量较差。基于规则的方法通过专家们共同制定的规则可以获得高质量的知识, 但是规则都是针对特定领域制定的, 灵活性较差。

新词发现也不外乎这两种方法。统计方法通过对词共现进行概率统计而实现的。这种方法适用于任何领域, 但是它们需要大量的训练语料。Lai和Wu^[3]基于统计方法对PLU(Phrase-Like Unit)的概率进行计算, 进而发现新词/短语, 结果证明基于新短语的分类方法达到了很好的性能。基于规则的方法是通过标注词典以及组词规则来识别新词, 但是该方法的困难是词性的歧义性和语法的灵活性, 另外词典中不可能包含所有的中文词, 也不能穷尽所有的组词规则。

本文在分析前人研究成果基础上, 本着从实用角度出发, 研究二者的融合方法实现新概念的抽取: 以快速的统计方法为工具, 自动获取特定领域的新词语、新概念; 在此基础上通过一系列的规则进行过滤。这样既吸收统计方法的快速, 又可保留专家系统方法的质量。

2 系统结构

基于概率统计和规则方法的新词发现系统结合了两种方法的优点, 能够快速且高效地在大量的文本中发现高质量的概念。图1系统地显示了概念发现的工作流程: 首先通过网络蜘蛛从Internet下载自己需要的语料, 然后对话料库进行HTML解析和分词处理, 并将文本表示成方便用N元语法进行统计的格式, 进而利用公式进行统计选词。此时的结果因统计方法自身的局限显得不是很理想。我们结合自然语言处理的方法, 进一步分析了统计选词后的数据, 总结出一些规律并且将这些规律表示成规则。这时就可以有选择地利用这些规则选词。当然, 规则选词后的结果依然有不尽如人意的词条, 但是这时的词条数量较少(大约3 000条), 这样就使得人工干预成为可能。经过人工挑选后的结果就是我们最终所得到的概念, 包括新词、词组和短语。

总的来看, 该系统由下列模块组成: 语料库的获取模块, 数据预处理模块, 统计模型的分析 and 实现模块, 统计选词模块, 规则选词模块。下面章节将会对各个模块进行详细的介绍, 特别是统计模型和规则选词模块。

基金项目: 国家自然科学基金资助项目(60173017, 90104021); 北京市自然科学基金资助项目(4011003)

作者简介: 贾自艳(1971—), 女, 博士生, 研究方向: 数据挖掘, 智能信息处理, 信息检索等; 史忠植, 研究员、博导

收稿日期: 2003-08-29 E-mail: jiazy@ics.ict.ac.cn

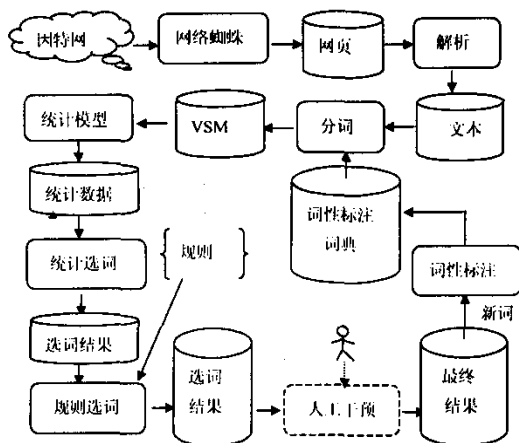


图1 概念发现系统的体系结构

3 统计模型

N-gram(N元语法)是计算语言学中经常使用的统计模型,也是概念抽取中经常用到的。是否N的值越大越好?其实不然。Lewis^[4]对文本分类中使用词组的效果进行了测试,研究表明以n-grams作为文本向量空间模型的补充可以提高分类性能,但是 $N>3$ 时性能不变,甚至降低。

同时有关统计信息显示,汉语中相邻两个词语组合成新词语的可能性是很大的,尤其是具体到某一特定领域时,特定概念通常都是由多个词语组合来表达的,其中两个词语的组合最为普遍。而且,两个词语组合又是多个词语组合的基础,因此本文以二元语法(bi-gram)作为统计模型。“二元语法”,就是只考虑相邻的两个词语所得到的语法和数据信息。

定义1 “二元语法”的统计模型设计如下:统计的基本元素是系统所用的通用分词词典中的词。统计的对象是限于任何两个标点符号之间的连续词序列,可以表示为“ $w_1 w_2 w_3 \dots w_n$ ”,从序列的第一个词语开始,依次记录相邻两个词语组合“ $w_j w_{j+1}$ ”($1 \leq j \leq n-1$)的共现串、共现文档名称、共现位置等信息,同时实现共现次数、共现文档数目的累计。这样就得到两词集合(bi-word),记为 $bi-word_i = \{w_j w_{j+1} | 1 \leq j \leq n-1\}$ 。为了便于描述,我们称第一个词“ w_j ”为首词,记为Fword;第二个词“ w_{j+1} ”为尾词,记为LWord。

3.1 数据预处理

由于该统计模型的计算量非常大,在统计之前需要经过数据预处理。预处理操作包括将HTML网页解析成文本的过程和将文本利用词性标注词典进行分词得到向量空间模型的过程。

网页数据是一种半结构化的数据,也就是说,组成网页文档的所有段落都与HTML标记相联系,这些HTML标记表明这些段落信息的重要性。我们利用网页解析器对网页进行解析,仅保留有用的文本信息,这样既不影响新词发现的质量,同时又能提高算法的效率。

接下来的工作是对这些文本语料库进行汉语自动分词。为了便于利用语法、语义信息形成知识和规则,我们为通用分词词典引入了词性分类,得到词性标注过的词典。该词典中含有近6万词条。这样就可以在对文本语料库自动分词的同时实现自动词性标注。当然词性标注中兼类词问题是很难解决的,考虑到我们的目的不是使词性标注达到百分之百的

正确,而是将其作为一种辅助信息,帮助我们分析二元语法统计结果的,所以把兼类词统一标成最为常用的词性。例如:“研究”一词在词典中既有名词词性也有动词词性,但是在文本标注过程中仅将其标记为名词。我们采用的分词算法是正向最大匹配法。预处理后的数据表示为: $DOC = \{(Ele_i, Tag_i, Loc_i) | \text{if } Ele_i \in DICT \text{ then } Tag_i \in TagS \text{ else } Tag_i = "0000", i = 1, 2, \dots, s\}$ 。其中DICT为词典,TagS词性标记符号集。 $s = \|DOC\|$ 为文档中元素的数目。

3.2 统计算法

为了便于数据的统计,将为每个词生成一个倒排文件,表示为: $word \rightarrow \{(Doc_1, Loc_1), (Doc_2, Loc_2), \dots, (Doc_m, Loc_m)\}$ 。其中每个二元组代表该词语出现过一次,出现的文档名称是 Doc_k ,该词出现的位置是 Loc_k ($1 \leq k \leq m$,设该词语在所有文档组成的语料库中出现了m次)。该词出现的文档频次为词语倒排文件中不同文档的数目。然后根据这些倒排文档数据对每个词根据二元语法模型统计其共现数据,统计算法为:

```
(1) for all InvertedFile(word)do
(2) FWord ← word
(3) for all pairs of (Doc,Loc) in InvertedFile(word) do
(4) Ele ← FindWord(Doc, Loc+1)
    //以文档名称为入口找到该文档的符号序列文件,
    //并读出其中位置为location+1的符号。
(5) if (Ele is not Chinese word) then do nothing
    //如果是标点或数字、英文等字符,则不予处理。
    else if (Ele did not occur)
        //如果发现该词没有出现过
        then record { LWord ← Ele; //共现辅词
                    Freq(Ele) ← 1; //共现次数1
                    Doc(Ele) ← Doc; //共现文档名
                    }
    else record { Freq(Ele) ← Freq(Ele)+1;
                  Doc(Ele) ← Doc; }
    //共现次数加1,同时记录文档名。
(6) endfor;
(7) endfor;
```

通过执行上述算法,得到了所有二元词语共现的信息: $FWord \rightarrow \{(LWord_i, Doc_i, Freq_i), \dots, (LWord_n, Doc_n, Freq_n)\}$ 。这样,我们就保存了大量的信息,以所有的FWord为主链,每个FWord都记录了在语料库中的出现总频次和文档频次,对该FWord所出现过的每一篇文档都记录了文档名称和该文档中的出现频次。每个FWord后面都跟了一系列与其共现的LWord,对每一个LWord,分别记录其与FWord共现的总频次和文档频次,以及每一篇文档的名称和在该文档中的共现频次,这样非常有利于我们进一步分析词语的构词能力。

3.3 统计选词

在实验中我们尝试了多种策略进行统计选词,但是因为篇幅所限,这里不逐一介绍,仅对最终采用的效果较好的方法进行介绍。

定义2 在文本上下文序列“ $w_1 w_2 w_3 \dots w_n$ ”中,若 w_i 后面紧跟词 w_j ,记为: $w_i \Rightarrow w_j$ 。 $w_i w_j$ 组成新词(短语)的可信度定义为 w_i 后出现词 w_j 的概率: $p(w_j | w_i) = df_{ij} / df_i$ 。其中, df_{ij} 代表词 w_i 和词 w_j 共现频次, df_i 代表词 w_i 出现频次。针对每个首词 w_i 分别计算与其共现的所有尾词 w_j (设共有K个)的共现频次均值:

$$E(df_i) = \sum_{j=1}^K p(w_j | w_i) \times df_j$$

这里定义统计选词原则为“共现频次在均值之上的词汇组合是好的”，即 $df_{ij} > E(df_i)$ 。

我们对计算机语料库进行了实验。最初的二元语法统计结果共有50多万共现词汇，通过该公式的初步筛选，挑出了24万余条共现词汇。从结果分析来看，其中很多都是好的，但不好的也不少，效率不是十分高。究其原因如下：(1)统计公式的制定是把所有的词语都统一对待，只考虑它们的共性，没有考虑它们在使用过程中的种种个性，而这些个性往往是它们形成新词的决定性条件。(2)语料库本身的选取不能很好地满足上面理想化公式的使用条件。(3)汉语自身的灵活性。

总之，单纯的统计方法对于语料库的选择和统计公式的制订有极大的依赖性，难以达到很高的准确度。必须增加必要的知识，通过规则的方法来提高准确度。

4 规则选词

在统计结果的基础上，加入适当的知识和规则，可以更为详尽地描述汉语词语在构词和使用过程中的个性特点，从而能够提高单纯基于统计的二元语法选词效果。通过大量语料分析，发现首词和尾词皆为单字形成新词的概率很大，多为社会上流行的新词新语。但其成词的规律与首词和尾词中至少有一个是多字词的组合很不相同，因此把它们单独取出来进行处理。

4.1 单字组合词规则

属于某些词性的词(见表1)本身不具有实际的概念意义，其功能主要是用来帮助造句的，这些词很少用来组成新词、新概念。

表1 不可组词的单字词性规则

词性	标记	例子
数词	U***	“一”、“二”、“十”、“千”
代词	P***	“他”、“它”、“你”、“我”
介词	R***	“在”、“于”、“从”
助词	Z***	“的”、“么”、“哉”、“啊”
象声词	S***	“叭”、“嘿”、“砰”
姓氏单字	NSUR	“王”、“李”、“赵”、“贾”、“徐”

除了表1中列出词性以外，其它词性的某些单字由于自身意义的原因，也很少用来组成新词新语，这里我们归为禁用词规则(见表2)。单字组合词汇分离出以后，依次经过下面的流程可得到最后的结果：不可扩展的单字列表过滤→只做主词的单字列表过滤→只做辅词的单字列表过滤→词性规则过滤→人工挑选。

表2 单字组合词禁用词规则

规则	例子	数量
不可扩展的单字	“沪”、“斯”等以及三级字库中的偏僻字	500
只做首词的单字	“上”、“前”等	14
只做尾词的单字	“除”、“加”、“至”等。量词(QOTH)和连词(COTH)	60

4.2 多字组合词规则

下面主要对首词和尾词中至少有一个是多字词的组合同义进行了分析。这里的部分方法和单字组合词类似，也可以

按照其组成新词的能力以及自身的特征分成以下4种类别的禁用词(见表3)。

表3 多字组合词禁用词规则

规则	特点	组成/例子	数量
禁用虚词	没有组词能力，只有构造句子能力，同时自身是没什么实际意义词	助词、连词、疑问词、感叹词等虚词	1 047
禁用实词	自身意义相当完整，几乎没有必要再组词的词	成语、俗语、叠词，多为形容词、副词、动词等	5 675
只做首词	可用来扩展新词，但通常只做首词	“最高”、“依次”、“中和”等	77
只做尾词	可用来扩展新词，但通常只做尾词	“专业”、“折扣”、“障碍”等	414

表3中的4类词的集合构成了筛选词的知识库，一旦发现禁用词或者自身意义相当完整，就不再继续进行共现统计。而对于只做首词和只做尾词的词也应当分别做出相应的处理。多字组合中存在两种比较特殊的名词，它们的成词率很高，这就是后缀和前缀类名词词汇。董振东先生的知网(<http://www.how-net.com>)特意分出了前缀词类和后缀词类两种类别，可见这两类词的作用是比较突出的。但是董先生所定义的词比较少，通过分析大量统计结果，总共形成了60个前缀名词，180个后缀名词。我们还发现能够与这两类词组合的词汇类别较多，因此将所有的前缀和后缀组合词汇分离出来专门处理。表4总结了多字组合的规则。

表4 多字组合规则

规则	成词率	规则	成词率
单字+多字	低	名词+副词	低
名词+名词	较高	职位/职称+姓氏	低
名词+动词	低	前缀名词+词	高
名词+形容词	低	词+后缀名词	高

总体来讲，从形成有意义的新概念的角度来说，还是名词和名词的组合同义成功率更大一些，其它词性的词汇组合成功率很小，为了保险起见，把所有可能是名词的兼类词都当作名词。现在，将以上规则形成相应的规则文件，然后编程实现。整个处理流程为：

禁用词过滤→后缀过滤→前缀过滤→名词过滤→特殊语义类过滤→首词是单字的过滤→不可扩展的实义词过滤→只做首词的词语列表过滤→只做尾词的词语列表过滤→人工挑选。

令人欣慰的是通过规则的层层过滤，可以获得较高质量的结果，大大减少人工干预的工作量。而且，上述的过滤知识文件都是与程序分离的，在系统执行过程中随时可能挖掘出新的规则，用以完善规则知识库，从而加强系统的智能化进而减少工作量。

5 实验结果

我们使用新词发现系统对计算机世界语料库(40MB)和《人民日报》财经新闻语料库(20MB)分别做了试验。图2显示了单字组合词的发现过程以及结果，图3显示了多字组合词的过滤情况。它们的发现过程是根据前面介绍的统计模型和过滤(选词)规则逐步完成的，每步执行结果在图中清楚地显示出来。从两图中可以看出我们制定的过滤规则是很有效

(下转第83页)

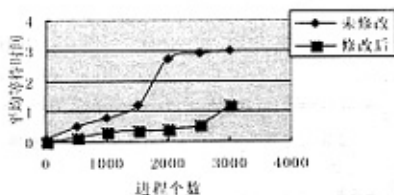


图4 平均等待时间对比

从中可以看到,修改后的Linux系统具有比原来更短的平均周转时间和平均响应时间,从而说明了在大用户量的系统中,多级反馈队列调度策略确实要优于Linux原来的基于时间片轮转的动态优先级调度策略。

6 总结与展望

Linux是一个开放源码、有广泛应用前景的多用途操作系统。作为有希望替代Unix的服务器操作系统,其还有一些性能上的问题亟需克服。本文在研究目前已有的普通Linux进程调度策略的基础上,提出了用多级反馈队列调度策略替

(上接第21页)

的:从最初大量的原始统计数据数据(计算机领域:50万条,经济领域:37万条),经过层层过滤后得到最终的结果(统计数据见表5)。从表5可以看出该系统能够帮助我们发现有益的新词、新概念。考虑到分词的效率和歧义性,我们只将通用的新词可以加入通用分词词典中,而其它的新词可加入领域词典中。

表5 新闻发现结果示例

领域	计算机领域		经济领域	
类型	单字组词	多字组词	单字组词	多字组词
数量	3 000	3 010	1 586	3 550
例子	“死机”、 “按钮”、 “字节”、 “字段”、 “总机”、 “声卡”、 “站点”、 “页眉”、 “鼠标”、 “网吧”、 “黄页”、 “师姐”等	“即插即用”、 “北大方正”、 “批处理”、 “结束符”、 “硬拷贝”、 “局域网”、 “差错率”、 “制造业”、 “神经计算模 型”、“演示 版”、“计算机 网络”等	“下岗”、 “树叶”、 “影碟”、 “造假”、 “韩国”、 “招标”、 “中标”、 “转账”、 “征婚”、 “诊所”、 “执着”、 “庄家”等	“安全感”、 “八达岭长 城”、“领导 班子”、“包 装箱”、“保 健品”、“保 障体制”、 “项目经理”、 “个体户”、 “菜篮子”、 “侏罗纪公园” 等

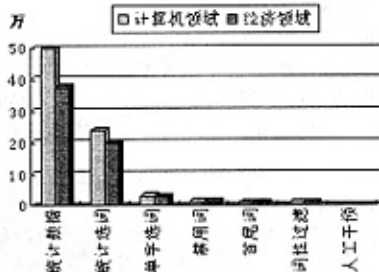


图2 单字词过滤过程结果显示

万方数据

代Linux现有的普通进程调度策略,从而使Linux系统具备良好的大用户量时的分时响应性能。这种多级反馈队列调度的策略虽然比Linux原有的进程调度策略复杂,但是可以有效地提高Linux服务器的性能,从而减小与高端Unix服务器的差距,为Linux更广泛的应用打下一个良好的基础。随着更多新的进程调度策略在Linux上实现,这种现代操作系统在服务器应用领域必将会有光辉的前景。

参考文献

- 1 Aivazian T. Linux Kernel 2.4 Internals. <http://www.linuxdoc.org>, 2002-08
- 2 Stallings W. Operating Systems: Internals and Design Principles (Fourth Edition). Prentice Hall, Inc., 2002
- 3 尤晋元. Unix高级编程技术. 上海: 上海科技文献出版社, 1994
- 4 毛德操, 胡希明. Linux内核源代码情景分析. 杭州: 浙江大学出版社, 2001
- 5 Torvalds L. Linux Kernel Source 2.4.17. <http://www.kernel.org>, 2002
- 6 屠 祁, 屠立得. 操作系统基础(第三版). 北京: 清华大学出版社, 2002
- 7 黄千平, 陈洛贤. 计算机操作系统. 北京: 科学出版社, 1989

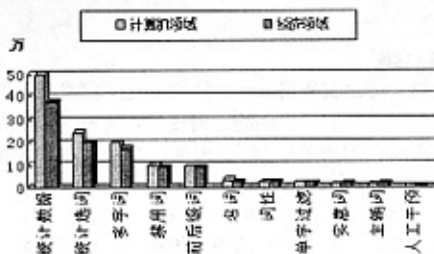


图3 多字词过滤过程结果显示

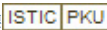
6 结束语

本文根据汉语的特点,采用基于概率统计和规则相结合的方法对新词的自动获取进行了有益尝试。这是一种适用于从大规模语料库中发现新知识的方法。该方法的不足之处是:使用统计公式进行统计选词的结果依赖于语料库的发散程度,而规则过滤的效率依赖于规则知识的获取和完备。下面的工作将尝试使用数据挖掘的一些算法(例如关联规则),以及Markov概率统计模型来提高系统的效率。

致谢 该研究工作得到首都信息发展有限公司的李蕾博士、郭祥昊博士、王楠和伏小妹等人的帮助和建议。这里表示深深谢意!

参考文献

- 1 Schutze H, Hull D, Pederson J.A Comparison of Classifiers and Document Representations for the Routing Problem. In Croft . (Eds.), Proceedings of SIGIR-95, 15th ACM International Conference on Research and Development in Information Retrieval, New York: ACM Press, 1995:229-237
- 2 Tan Chademeng, Wang Yuanfang, Lee Chando.The Use of Bigrams to Enhance Text Categorization. Information Processing and Management, 2002,38: 529-546
- 3 Lai Yusheng, Wu Chungshien.Meaningful Term Extraction and Discriminative Term Selection in Text Categorization via Unknown-word Methodology.ACM Transaction on Asian Language Information Processing, 2002,1:34-64
- 4 Lewis D.Feature Selection and Feature Extraction for Text Categorization. In Proceedings of a Workshop on Speech and Natural Language , San Mateo, CA: Morgan Kaufmann, 1992:212-217

作者: 贾自艳, 史忠植
作者单位: 贾自艳(中国科学院计算技术研究所, 北京, 100080; 中国科学院研究生院, 北京, 100039), 史忠植(中国科学院计算技术研究所, 北京, 100080)
刊名: 计算机工程 
英文刊名: COMPUTER ENGINEERING
年, 卷(期): 2004, 30 (20)
被引用次数: 13次

参考文献(4条)

1. Schutze H; Hull D; Pederson J A Comparison of Classifiers and Document Representations for the Routing Problem 1995
2. Tan Chademeng; Wang Yuanfang; Lee Chando The Use of Bigrams to Enhance Text Categorization [外文期刊] 2002 (4)
3. Lai Yusheng; Wu Chunghsien Meaningful Term Extraction and Discriminative Term Selection in Text Categorization via Unknownword Methodology 2002
4. Lewis D Feature Selection and Feature Extraction for Text Categorization 1992

本文读者也读过(5条)

1. 王立希, 王建东, 汪静, WANG Li-xi, WANG Jian-dong, WANG Jing 基于数据挖掘的新词发现 [期刊论文] - 计算机应用研究 2006, 23 (12)
2. 王文荣, 乔晓东, 朱礼军, Wang Wenrong, Qiao Xiaodong, Zhu Lijun 针对特定领域的新词发现和新技术发现 [期刊论文] - 现代图书情报技术 2008 (2)
3. 吴春颖, 王士同, 蔡崇超, WU Chun-ying, WANG Shi-tong, CAI Chong-chao 一种基于新词发现的Web文本表示方法 [期刊论文] - 计算机应用 2008, 28 (3)
4. 罗智勇, 宋柔, LUO Zhi-yong, SONG Rou 基于多特征的自适应新词识别 [期刊论文] - 北京工业大学学报 2007, 33 (7)
5. 林自芳, 蒋秀凤, LIN Zi-fang, JIANG Xiu-feng 基于词内部模式的新词识别 [期刊论文] - 计算机与现代化 2010 (11)

引证文献(13条)

1. 韩艳, 姚建民, 朱巧明, 张晶 不限领域的中文新词的识别研究 [期刊论文] - 郑州大学学报 (理学版) 2008 (3)
2. 汪青青 现代汉语新词特征探析 [期刊论文] - 现代语文 (语言研究) 2009 (7)
3. 吴春颖, 王士同, 蔡崇超 一种基于新词发现的Web文本表示方法 [期刊论文] - 计算机应用 2008 (3)
4. 徐艳山, 王国才 二维化信息的三维化加密算法研究 [期刊论文] - 计算机应用 2010 (4)
5. 王文荣, 乔晓东, 朱礼军 针对特定领域的新词发现和新技术发现 [期刊论文] - 现代图书情报技术 2008 (2)
6. 孔晨妍, 侯汉清 《中国图书馆分类法》类目更新途径之探讨 [期刊论文] - 图书馆工作与研究 2007 (1)
7. 周蕾, 朱巧明, 李培峰 一种基于统计和规则的未登录词识别方法 [期刊论文] - 南京大学学报 (自然科学版) 2005 (z1)
8. 周蕾, 朱巧明 基于统计和规则的未登录词识别方法研究 [期刊论文] - 计算机工程 2007 (8)
9. 刘华 一种快速获取领域新词语的新方法 [期刊论文] - 中文信息学报 2006 (5)
10. 黄轩, 李熔烽 博客语料的新词发现方法 [期刊论文] - 现代电子技术 2013 (2)
11. 韩艳, 林煜熙, 姚建民 基于统计信息的未登录词的扩展识别方法 [期刊论文] - 中文信息学报 2009 (3)

12. [周蕾](#) [中文未登录词识别的研究及在汉字输入法中的应用](#)[学位论文]硕士 2005
13. [杨江](#), [赵晗冰](#) [语言监测中的词语新义自动发现](#)[期刊论文]-[湖南科技大学学报（社会科学版）](#) 2013(1)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_jsjgc200420009.aspx