面向Internet的中文新词语检测*

邹纲· 刘洋· 刘群·
中科院计算技术研究所数字化实验室
北京 100080

孟遥· 于浩· 西野文人·
富士通研究开发中心有限公司
北京 100081

亢世勇·
烟台师范学院中文系
烟台 264025

摘要：

随着社会的飞速发展，新词语不断地在日常生活中涌现出来。搜集和整理这些新词语，是中文信息处理中的一个重要研究课题。本文提出了一种自动检测新词语的方法，通过大规模地分析从 Internet 上采集而来的网页，建立巨大的词和字串的集合，从中自动检测新词语，而后再根据构词规则对自动检测的结果进行进一步的过滤，最终抽取出采集语料中存在的新词语。根据该方法实现的系统，可以寻找不限长度和不限领域的新词语，目前正应用于《现代汉语新词语信息（电子）词典》的编纂，在实用中大大的减轻了人工查找新词语的负担。

关键词：中文处理 新词语 自动检测

中图分类号：TP391

Internet-oriented Chinese New Words Detection

ZouGang LiuYang LiuQun
Institute of Computing Technology,Chinese Academy of Sciences Beijing 100080

MengYao YuHao
Nishino Fumihito
Fujitsu Research & Development Center Co.,LTD Beijing 100081

KangShiyong·
Yantai Normal University
Chinese Department Yantai 264025

Abstract:

With the fast development of the society,more and more new words come out in our life. It is one of the important topics in Chinese natural language processing to collect those new words. A method is presented for detecting these new words automaitcally in this paper. Through analysing webpages grabbed from the Internet, a large word and string set is built, which new words are detected from and filtered by rules. At last new words which exist in the webpages grabbed are extracted. The system built in this way can find new words in any length and in any field.Now it is applying to the compilation of Modern Chinese New Word Information Dictionary. It reduced human labor a lot in practise.

KeyWords: Chinese Language Processing New Word Automatic Detection

*作者简介： 邹纲 ：男，1978 年生，硕士研究生，研究方向是自然语言处理
刘洋 ：男，1979 年生，硕士研究生，研究方向是统计机器翻译，机器翻译自动评测
刘群 ：男，1966 年生，副研究员，主要研究方向为机器翻译和自然语言处理。

1 前言

自然语言中新词语的不断涌现是一个客观规律。随着经济、社会的飞速发展和对外交流的日渐频繁，特别是Internet的普遍使用，这一现象变得更加明显。据中国语言文字工作委员会专家曾做的一个比较保守的统计，中国自改革开放的 20 年来平均每年产生 800 多个新词语[1]。

由于汉语中词语的定义的模糊性，很难给出一个新词语的确切的定义，在现有研究的基础上，我们认为对于新词语可以从下面两个方面把握：(1)从词典参照的角度来说，新词语是指通过各种途径产生的、具有基本词汇所没有的新形式、新意义或新用法的词语[2]。新词语的特点在于“新”，这个“新”具体表现在词形、词义和词语的用法上。鉴定新词语的参照系是现代汉语基本词汇的词形、词义和词语的用法。着眼于一个词语的词形、词义和用法，将其与现代汉语基本词汇的词形、词义或用法进行比较，只要在这三个方面的任何一点上不同，就认为它是新词语。基本词汇的代表是《现代汉语词典》和《汉语大词典》的主体词汇。(2)从时间参照角度来说，新词语是出现在某一时间段内或自某一时间点以来所首次出现的具有新词形，新词义或者新用法的词汇[3]。新，就体现了与时间相关的特点。比如可以把改革开放以来出现的词语如“经济特区”,“下海”,“打工”等等称为新词语。

从大体上说，从语言学角度，汉语中的新词语按照来源可以分为以下几类[2]：

(1) 命名实体：包括人名、地名、商品名、公司字号、机构名等；
(2) 缩略语：如“非典”、“计生委”等；
(3) 方言词：如“靓”、“埋单”等；
(4) 新造词：如“伊妹儿”、“美眉”等；
(5) 专业术语：如“非典型肺炎”、“蓝光光盘”等；
(6) 音译词：如“酷”、“秀”、“克隆”等；
(7) 字母词：如 WTO、APEC 等；
(8) 词义、用法发生变化的旧有词语：如“下课”、“充电”等。其中还包括一种“旧词新用”的语言现象，比如“高就”、“赏光”等，很长时间不用了，最近又重新出现在语言中。

就目前而言，新词语自动检测的困难主要在于：

(1) 汉语的词与词之间没有间隔；
(2) 除了命名实体和字母词外，其他具有新词形的词语的构成基本上没有一个比较普遍的规律；
(3) 对于低频的具有新词形的词语识别比较困难；
(4) 对于词义、用法发生变化的旧有词语的检测更加困难。

现有的新词语自动检测的研究，以命名实体类居多，在汉语的命名实体识别研究中，又以人名、地名、音译名识别率较高，准确率和召回率都可以达到 90％以上[4]，机构名构成规律较为复杂，识别准确率和召回率较低一些。其他类型的命名实体研究很少，识别率更低。至于其他类型的新词语自动检测的研究就更少。由于命名实体的识别研究已经比较充分，因此本文目的在于研究非命名实体类的具有新词形的词语的自动检测。为方便起见，下文中提到的新词语均指非命名实体类的具有新词形的词语。

非命名实体类的具有新词形的词语的自动检测的研究，目前国内主要的研究方法有两种：一是规则的方法，通过建立专业词库，模式库和规则库，对语料进行识别处理[5]。二是统计的方法。利用重复串的信息，提取高频的串，然后再利用语言知识排除不是新

词语的垃圾串[6][7][8][9][10]，或者是计算相关度，寻找相关度最大的字与字的组合[11]。规则的方法主要缺点在于局限于某个领域，并且需要建立规则库等等。统计的方法，一般都是限于查找二字，三字和四字的新词语。

本文提出的新词语自动检测方法，目标是大规模的处理 Internet 网页，从一个整体的角度，寻找某一时间点后首次出现的不限领域和长度的任意新词语。

根据本文提出的方法实现的系统，在实验中，准确率在 30％至 40％间，召回率在 90%左右，已应用于《现代汉语新词语信息（电子）词典》的编纂上，大大的减轻了人工查找新词语的负担。

2 面向 Internet 的中文新词语的检测

2.1 基本思想

新词语，就是已有汉字或词语的一种组合。而一种组合之所以能固定下来，被认为是一个新词语，而不是一种临时组合的词组，该词语必定要经常性重复出现。不仅要在一篇文章中多次出现，而且要在很多文档中反复出现，这是一个新词语被承认的必要条件。新词语另外一个规律是具有时效性。一般而言，是在某个时间点之前新词语几乎没有出现过，而从某一个时间点开始出现，越来越频繁，而后稳定下来，比如改革开放以来出现的“经济特区”等词，有的甚至会渐渐消亡，比如“粮票”、“布票”等这些建国以来出现的新词语。

因此从一个整体的角度看，新词语有两个特征：

（1） 具有重复出现的规律；

（2） 具有时间规律，即新词语总是在某个时间点之后出现并且流行。

此外在词法分析中，绝大部分的新词语都是未登录词而被切成散串。比如专业术语“非典型肺炎”，在词法分析中被切成了“非 典型 肺炎”，缩略语“非典”被切分成“非 典”。如果这些词语在一篇文章中重复出现两次以上，那么通过重复串查找就可以将它们找回。

综合上述的这些特征，本文提出的新词语自动检测的基本思想是：首先大规模处理网页，对于切分后的网页内容，用重复串查找寻找新词语，但同时也不可避免的找到了一些固定搭配和噪音串。接着根据给定的时间，建立一个给定时间之前的大规模的词与串的背景词串集合，这个集合里面不仅包括了大部分已有的词语，并且还包括了噪音和固定搭配。然后我们在这个背景词串集合的基础上，通过评价函数对于给定时间以后的词和串进行比较和评价，从中得到新词语候选。最后用过滤规则对新词语候选进行过滤，得到最终的新词语结果。具体的流程如图 1 所示。

2.2 重复串查找

重复串查找的目标是在一段文本中将所有出现在指定次数以上的重复串都查找出来。查找可以以字为单位，也可以以词为单位。本文采用的是基于词的重复串查找的方法。与不分词进行基于字的重复串查找的方法相比，分词后基于词的重复串查找的优点在于可以大大减少垃圾串产生。比如对于“…..中国人民…..中国人…..”这段话，不分词的话，基于字的重复串查找将会将“中国人”作为一个重复串查找出来。但是如果分词后进行基于词的重复串查找的话，则“… 中国 人民…. 中国 人 …..”这段话中，“中国人”不将作为一个重复串被找出来。

由于需要大规模的处理十万至百万级的网页，因此要求重复串查找算法必须快速。针对这些要求，并且对比了其他重复串查找的算法后，我们提出了我们的重复串查找算法。该算法的时间复杂度的上界是 O(kn)，其中 k 是文本中最长的重复串长度，空间复杂度是线性的。由于一般自然语言文本中，常见的重复是不长的词和短语，因此在实用中，完全能满足本文提出的要求，但是限于本文的篇幅，这儿不做算法的描述以及时间空间复杂度的证明。

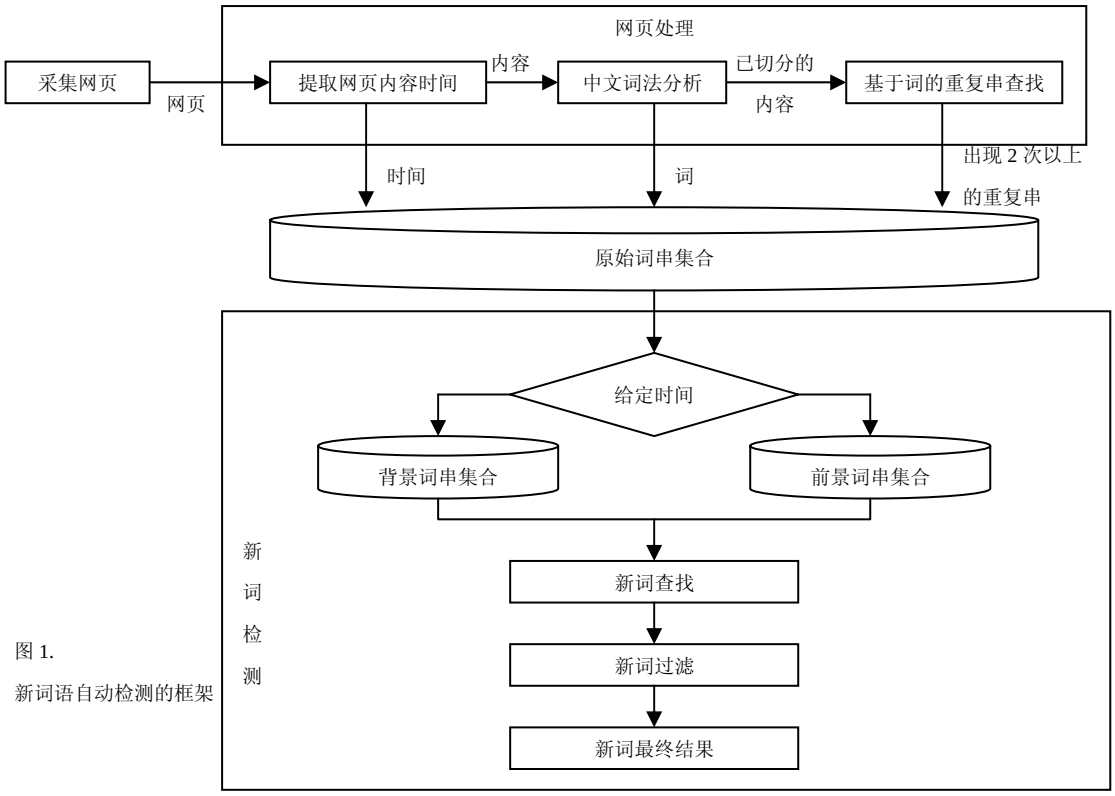

图 1.

新词语自动检测的框架

2.3 新词语查找

网页处理结束后，生成了一个巨大的词与串的原始信息集合。新词语的查找是在这个集合的基础上进行。原始信息集合由两张表组成，其例子如图 2 所示。图 2 的左表存放网页的原始信息，右表存放每篇网页中出现 2 次或者 2 次以上的所有的词和串。

页面 id	路径名	日期
1083	\774\97632.html	2003-06-11
1084	\774\97642.html	2003-06-11
1085	\774\97643.html	2003-06-11
1086	\774\97644.html	2003-06-11
…	…	…

对应左表的页面 id	词串	词性	频度
1083	条件	n	2
1083	任务	n	2
1083	非 典	n ng	14
1083	抗击 非 典	v n ng	2
1083	非 典 疫情	n ng n	3
1083	非 典 患者	n ng n	3
…	…	…	…

图 2 原始信息集合示例

从原始信息集合中生成背景词串集合的过程，是将指定日期之前的所有文档的词或者串的总频度和文档频数统计出来。与此类似，将指定日期之后的所有文档的词或者串的总频度和文档频数统计出来，即成为前景词串集合。这儿的文档频数指的是该词或串在多少篇网页中出现过。背景词串集合的例子如图 3 所示。

假设用 S_b 来表示背景词串集合，用 S_f 来表示前景词串集合，其中 $S_b=\{e|e=(word,totalfreq,docfreq)\}$，$S_f$也类似的表示。则评价函数可以表示成如下

$$f(x)=\begin{cases}1 & x.totalfreq \geq m \,\&\, x.docfreq \geq n \,\&\, \\ & \forall z(z \in S_b -> z.word \neq x.word) \\ 0 & \text{其他情况}\end{cases}$$

其中$x \in S_f$，m和n是预设的阈值。输出 1 表示该词是新词候选。

词或者串	词性序列	总词频	文档频数
保护伞	n	74	16
生态 农业	n n	74	24
法人 治理 结构	n vn n	74	18
形式主义	n	74	14
外国 专家	n n	74	14
困难 群体	an n	74	29
农村 义务教育	n nl	74	22
政务 公开	n vn	74	17
刑讯 逼供	n vi	74	8
…	…	…	…

图 3 背景词串集合示例

2.4 新词语过滤

由于背景词串集合不可能将所有的词与词之间的常用搭配都收录进去，因此在新词语查找生成的新词语候选中，存在着很多非新词语的垃圾串。为了尽可能的除去这些垃圾串，结合新词语的词形来考虑构成新词语的词性序列。下列的词性标注标准采用的是计算所汉语词性标记集。

在词法分析的过程中，新词语一般被切成散串，表现成三种情况：单字词之间的组合，单字词和多字词的组合以及多字词之间的组合。通过对大量的新词语词性序列的观察和分析，我们发现由于单字词词性兼类的现象比较严重，因此单字词的词性标注不一定准确。比如：边/d 警/ng，分/qt 尸/ng 案/ng，尤其是人名和地名以及缩略语，词性标注更是不准确，例如：库/n 赖/v，淤/vg 溪/ng 镇/n，秦/tg 碧/ng 天/n，临/v 管/v 会/v。而多字词的词性基本准确，比如：斩首/v 行动/vn，补偿/vn 指导/vn 单价/n，多边/b 会谈/vn，高考/vn 移民/n。

此外根据观察，新词语候选中的垃圾串，一部分是日期和数字相关的串，比如有关股指的串“1000 点”和日期“2003 年 11 月 6 日”等，另外有一部分垃圾串词性不符合构词规则，比如：“岁/qt 就读/vi 于/p”等。

通过对上面的新词语和垃圾串的特点的分析，可见可以通过构词规则来过滤掉一部分的确定不是新词语的垃圾串。此外由于寻找的新词语不限长度，因此用正则表达式来表示过滤规则比较方便和高效。参照文献[6]，我们从垃圾串中总结出的一部分过滤规则如下：

1. “[a-z]*d” 表示所有以副词结尾的词性序列；
2. “u[a-z]*” 表示所有以助词开头的词性序列；
3. “[a-z]*u” 表示所有以助词结尾的词性序列；
4. “[m]+” 表示单纯的数字；
5. “[t]+” 表示所有的日期；
6. “[a-z]*c[a-z]*” 表示所有包含连词的词性序列。
7. “q[a-z]*” 表示所有以量词起始的词性序列。

此外，由于单字词的词性标注不是很准确，所以在被过滤掉的垃圾串中，还要将非日期和数字的单字词召回。

2.5 优缺点讨论

从理论上分析，这种做法的优缺点在于

[1] 在背景数据库的支持下，可以过滤掉大多数的垃圾串，可以过滤掉已有的词。因为背景数据库不仅存入了已有的词，还存入了很多固定搭配及偶然噪音串等等。

[2] 可以寻找前言中所述的“旧词新用”这种新词。因为“旧词新用”的词的频度从时间上看，必然是在最近很长一段时间内很少出现的，然后在一段时间内出现的频度增加。

[3] 由于重复串查找算法寻找的串不限长度，因此存入数据库的串都是不限长度和领域的，从而这种方法可以寻找某个给定日期以后出现的不限长度和领域的新词语。

[4] 由于在每篇文章中查找重复串的阈值是 2，因此绝大部分高频的新词语可以找到，召回率比较高。

这么做的不足是：

[1] 由于在每一篇网页中，我们只能将出现频度大于或者等于 2 的串都提出来，因此那些在所出现的文章里频度都是 1 次的新词语不会被找到。但我们认为，新词语在所有其所在的文章中都只出现一次的情况比较少，大部分的新词语既然是被人们所承认和流行，那么必然会在文章中反复出现，因此对于结果影响比较小。

[2] 因为在原始信息数据库中存入的词或者串都是在文章中出现 2 次或者 2 次以上的词或者串，由于那些词或者串还可能在其他文章中出现 1 次，因此从原始信息数据库中统计出来的词或者串的总频度和文档频数并不是该词或者串的准确的总频度和文档频数。因此对于评价函数中的分别代表总频度阈值的 m 和文档频数阈值的 n 来说，很容易过滤掉这样的新词语：它们只在少于 n 篇的文章中出现 2 次或以上，在其他很多的文章中都只出现过一次。这是我们以后需要加以改进的地方。就目前而言，可以设低阈值 m 和 n。

3 系统实现与分析

3.1 系统实现和实验

实现的整个系统中，网页采集利用了一个共享的网页采集软件，中文词法分析则采用了计算所开发的汉语词法分析系统 ICTCLAS，词和重复串的存储用了 mysql 数据库。

由于人民日报系的报纸一般用语比较规范和严谨，一般而言，上面出现的新词语都是大家所认可的，因此从人民网采集了人民日报系的十三份报纸，一共有 647684 张新闻网页，时间跨度从 2000 年 1 月至 2003 年 12 月。

在网页预处理阶段，我们从新闻网页上提取了新闻的内容和新闻的时间，接着进行分词和重复串查找，然后以 2003 年 1 月 1 日为界，首先将这十三份报纸建成一个背景词串数据库，里面包含了 1820904 条词和串。然后在这个背景词串集合的基础上，经验的设定评价函数阈值 m 为 8，阈值 n 为 4，选择人民日报华东新闻和江南时报，以 2003 年 1 月 1 日为界将它们建成了两个前景词串集合，进行了以下两次实验：

报纸名称	人民日报华东新闻	江南时报
网页数量（2003 年 1 月至 12 月）	6295	41948
前景词串集合中词串的条数	23696	129930
新词语候选数目	82	605
新词语数目	30	146
未加过滤的准确率	36.6%	24.1%
未加过滤的召回率（大约）	100%	100%
未加过滤的 F 值	53.6%	38.8%
过滤后的准确率	40.30%	32.02%
过滤后的召回率	93.3%	93.8%
过滤后的 F 值	56.3%	47.7%

其中：准确率＝结果中正确的新词语数目/新词语候选数目×100%；

召回率＝结果中正确的新词语数目/前景词串集合中存在的新词语数目×100%。

需要说明两点，一是由于用来实验的集合很大，要是人工从人民日报华东新闻的 6295

份或者江南时报的41948份网页中寻找出所有的2003年度首次出现的新词语作为标准答案，是很费时费力的。而考虑到重复串查找的时候是将每篇文章中重复频度大于1的串都提取了出来，所以只要新词语在某篇文章中重复出现两次以上，就可以将其找到，并且我们认为在人民日报华东新闻或者江南时报中，在其所出现的文章里都只出现一次的新词语极少，因此我们把m=0、n=0时的新词语结果经过人工确定后作为新词语的标准答案，然后进行上述实验。二是，本文上述指标都是对于某个时间点首次出现的新词语进行考察的，而不是一份报纸中所有的新词语。

由实验可见，引入自动过滤后，F值提高了，整个系统的性能有所改进。

此外为了对比背景词串集合大小对于新词语查找的准确率的影响，我们选取了两份报纸，重新建立背景词串集合，进行了下面的实验：

背景词串集合大小	1820904(十三份报纸)	524684（两份报纸）
人民日报华东新闻新词语准确率 （未加过滤,2003/1-2003/12）	36.6％	6.6％
江南时报新词语准确率 （未加过滤, 2003/1-2003/12）	24.1％	12.0％

由实验可见，背景词串集合越大，越能提高新词语查找的准确率，这说明了背景词串集合过滤掉了大部分的噪音和固定搭配。

3.2 实验结果

下面是从两次实验中找到的2003年度的部分新词语：

一字新词语：

SARS　CEPA　THG　IVF

二字新词语：

非典　边警　几比　抗非　留观　留验　司考　淘碟　神五　性商　疑似　整人　迅驰……

三字新词语：

12强赛　百富榜　报检员　定销房　公投法　柜柜通　蓝筹股　临管会　国资委　漂亮MM　色老总　特巡警　重洽会　面贴膜　非典办…….

四字新词语：

笔形手机　第三商圈　多边会谈　高考移民　六方会谈　末代甲A　绿色巨人　绿化会战　世遗大会　物管主任　公推公选　美男作家　美女棋手　两个率先　产业同构　拆迁新政…….

四字以上新词语：

非典型肺炎　本息还款法　补偿指导单价　第三步兵师　国家渔政局　龙胆泻肝丸　突发公共卫生事件应急条例　重大项目投资洽谈会　新型农村合作医疗　全路面起重机……

分析实验结果，还发现：

1. 我们使用的词法分析软件ictclas具有命名实体识别的功能，因此在建立原始信息数据库之前，我们去掉了所有的词法分析中寻找出来的命名实体。但是在我们的新词语查找结果中，能召回部分机构名和人名。比如“国家渔政局”被切分成了“国家渔 政局”，因而在词法分析的下一步中未能将其识别成机构名，但是在新词语查找过程中被寻找出来了。又如“开心 大 药房”虽然切分正确，但是词法分析的命名实体识别模块没有将其找出，但是在新词语查找的结果中就出现了。
2. 查找出来的结果很多都是一些当年的热点问题。比如有关“和平解决朝核问题”，

“要求人民币升值”，“美英联军”等，这些短语出现的频率都相当的高。

3. 有的新词语伴随着一连串的短语出现。比如我们在查找到新词语“非典”的同时，新词语结果里同时出现了“非典病房”，“非典患者”，“非典疫苗”，“非典影响”，“感染非典” “非典病区”，“非典康复者”，“非典疑似病人”，“非典时期”等一系列与非典相关的短语。
4. 对于很多垃圾串，从词形或者词性角度来说难以加以判断和过滤。比如：“一/m 辆/q 坦克/n”和“三/m 个/q 代表/n”单从词性和词形上就无法区分。还比如短语“房价/n 涨幅/n”，从词性和词形上也无法区分是否是垃圾串。
5. 有些垃圾串是由于词法分析程序的错误造成的，比如：“是非典”被切成了“是非典”，还有比如“的的姐”被切分成了“的的 姐”。

从实验结果可见，本文所提的方法对于寻找各种类型的不限长度和领域的新词还是比较有效的

4 结论

汉语中新词语的不断涌现是一个客观规律。而随着 Internet 的普遍使用，这一现象变得更加明显。因而汉语新词语发现的研究具有重要的实际意义和现实需求。本文提出了一种自动检测新词语语的方法，通过大规模地分析从 Internet 上采集而来的网页，建立巨大的词和字串的集合，从中自动检测语，而后再根据构词规则对自动检测的结果进行进一步的过滤，最终抽取出采集语料中存在的新词语。对人民日报华东新闻和江南日报的初步实验结果显示：本系统的新词语发现精确率在 30%以上，而召回率可以达到 90%以上。实验表明，本文提出的方法可以发现观察文本中出现的大部分新词语。根据本文方法实现的系统，已经应用于《现代汉语新词语信息（电子）词典》词典的编纂，大大减轻了人工的负担。

5 今后的工作

新词语的选择和垃圾串的排除是影响系统性能的关键部分，针对目前准确率还不是很高的情况下，我们下一步将在背景词串集合的规模，新词语的选择标准以及垃圾串过滤等方面努力。首先是进一步扩大采集规模，由此建立了规模更加庞大的原始信息集合，其次是为采集的网页建立倒排表，以统计出每个词或者串的出现的真实的总频度和文档频数，第三是以一定的时间间隔从原始信息集合中统计生成这段时间间隔内的背景词串集合，从而可以追踪和绘制每个词或串出现的时空的曲线，以利于更准确的寻找新词。

致谢：

感谢张华平师兄仔细地阅读和修改了本文的初稿，并且提出了宝贵的意见。俞鸿魁同学对于本文的完成和新词语方面的研究也提出了很多有益的建议。在此一并致谢。此外非常感谢评审专家和编辑们对于本文提出的宝贵修改意见。

5 参考文献

[1] 张德鑫.水至清则无鱼－－我的新生词语规范观[J]. 北京大学学报（哲社版），2000，200005：p.106-119，http://www.hubce.edu.cn/cbb/qwjs/lib/33118.html

[2] 亢世勇等.《新词语大词典》前言

[3] 高永伟.英语国家对新词的研究[N].译者文苑，1999，http://www.cn-trans.com/cm-23.htm

[4] Hua-Ping ZHANG, Qun LIU.et al, Chinese Name Entity Recognition Using Role Model[J]. Special issue "Word Formation and Chinese Language processing" of the International Journal of Computational Linguistics and Chinese Language Processing, 2003, vol.8 No.2:p.29-60

[5] 郑家恒，杜永萍，宋礼鹏，农业病虫害词汇获取方法初探[A].孙茂松，陈群秀.语言计算与基于内容的文本处理[C].北京：清华大学出版社，2003，p.61-66

[6] 郑家恒，李文花.基于构词法的网络新词自动识别初探[J].山西大学学报（自然科学版），2002，25(2)：p.115-119

[7] 韩客松，王永成，陈桂林.无词典高频字串快速提取和统计算法研究[J].中文信息学报，2001，第 15 卷第 2 期：p.23-30

[8] 刘挺，吴岩，王开铸.串频统计和词形匹配相结合的汉语自动分词系统[J].中文信息学报，1998，第 12 卷第 1 期：p.17-25

[9] Craig G.Nevill-Manning, Ian H.Witten. Identifying Hierarchical Structure in Sequences:A linear-time algorithm[J]. Journal of Artificial Intelligence Research, 1997, 7:p.67–82

[10] 沈丽琴，施勤，柴海新.“自动新词提取方法和系统”，IBM 公司专利，，申请号：00126471.0

[11] 黄萱菁，吴立德，王文欣，et al.基于机器学习的无需人工编制词典的切词系统[J].模式识别与人工智能，1996，第 9 卷第 4 期：p.297-308