基于 WEB 的计算机领域新术语的自动检测*

刘知远　　孙茂松

清华大学计算机科学与技术系，清华信息科学与技术国家实验室，北京 100084

Email: liuliudong@gmail.com sms@tsinghua.edu.cn

摘　要：本文主要介绍“基于 WEB 的计算机领域新术语的自动检测”算法的设计和实现。随着计算机技术的迅猛发展，英语中每天都会出现大量的该领域的新术语，如何将这些新兴术语及时发现并纳入到汉语中来，是一个迫切而非常有意义的工作。该算法正是基于这一需求而设计实现的。其核心思想是，首先通过语料库的比对，找到“自某一时间点”以来在计算机类语料库中新出现的词语，即候选新术语，它们满足新术语的第一个特征；然后通过考查候选新术语在时间上的频度曲线，找到其中被广泛地应用，而非昙花一现的词语，确定为新术语。

关键词：自然语言处理，新术语，自动检测，统计，N 元词串，RSS

Web-Based Automatic Detection for IT New Terms

Liu Zhiyuan　　Sun Maosong

Department of Computer Science and Technology, Tsinghua University, Beijing 100084

Email: liuliudong@gmail.com sms@tsinghua.edu.cn

Abstract: This thesis introduces the algorithm “Web-Based Automatic Detection for IT New Terms”. With the rapid development of computer science and technology, a large number of new terms in the field are emerging. It is a very meaningful work to detect for these new terms in time and translate them into Chinese. The algorithm is designed based on the urgent needs. Generally new terms have two characteristics. First, new terms should emerge for the first time since some time. Secondly, new terms should be universally recognized and used widely. The algorithm detect for new terms based on their two characteristics. First of all, we compare different corpus built based on time to find candidates which meet the first feature. Then we adopt time series analysis to check the frequencies of these candidates for final new terms which meet the second feature.

Keywords: NLP，New Term，Automatic Detection，statistics，N-gram，RSS

1. 引言

随着社会与科学技术的飞速发展，日常语言中的新词语不断涌现。在英语世界，以收录新词语而闻名的Barnhart Dictionary of New English，其出版刊物The Barnhart Dictionary Companion称，每年可以提供约1,500个新词语和新语义[1]。在中国，语言文字工作委员会专家曾做的一个保守统计，中国自改革开放以来20年平均每年产生800多个新词语[2]。因此新词语在中国也成为研究的热点。汉语中的新词语，有相当大的比例来自英语。尤其是随着信息技术的飞速发展，英语世界每天都会产生大量的计算机领域特有的新术语，是汉语新词语的重要来源。总的来讲，计算机领域的新术语一般来自新技术、新思想、新产品。

正因为计算机领域的飞速发展以及该领域新术语的大量涌现，我国面临如何及时发现、引入和翻译这些新术语的问题，让汉语和中国的科技发展能够与世界IT产业发展保持同步。但是，目前缺少一个权威的机构及时地检测英语中新产生的术语并给予规范的命名，使得当前汉语中外来计算机类新术语的使用纷繁芜杂。因此及时发现计算机领域新术语并给出合理翻译的工作，是非常必要的。但这些新出现的术语散落在海量的网络文本中，靠人工去进行挖掘和检测是不可想

*本文承国家自然科学基金(项目号 60573187, 60621062 和 60520130299)的资助。

象的，因此亟需一个可以检测新闻网站并自动发现新术语的算法。而“基于WEB的计算机领域新术语的自动检测”算法正是为满足这一需求而进行的研究。

2. 前人工作

新词发现主要采用的方法不外乎基于统计和基于规则两种，或者这两种方法的综合。统计方法主要基于词内部的联系紧密度[3-6]，而规则方法则主要基于模板特征库、词语构造规则等[7-9]。在“基于WEB的计算机领域新术语的自动检测”的研究中，将主要使用基于统计的方法。英国利物浦大学将其研究发现的新词整理成名为“Neologisms in Journalistic Text”的网页[†]。他们从1989-2000年的报纸《Independent》和2003年的报纸《Guardian》上收集新词语。在收集过程中，他们使用了一种计算机过滤软件将新词识别出来，整理的新词表来看，全部是新词，而非词组，大多是新造词，少部分是旧词新意。而我们的研究，则是自动检测计算机领域新词语。也就是说不仅包括新词，还会包括新词组。这是因为在日新月异的信息科学的发展中，新技术的产生总会催生出新的描述词汇，这些词汇并不仅限于新词，有相当的部分是新词组，例如“Data Mining”，“Information Retrieval”等，它们多是旧词的新组合。因此，不仅新词，而且新词组，都是要求能够检测发现，也是本算法与利物浦大学研究项目的不同之处。

本文利用已有的自然语言处理技术研究设计“基于WEB的计算机领域新术语的自动检测”算法。

3. 算法描述

术语是某个领域的专用词语，它首先是词语。而新术语，首先应该具备新词语的特点。因此要对新术语进行自动检测，首先要明确新词语的涵义和特点。一般认为，新词语包括新词和新语。对于英语来讲，新词是单词，而新语是词组或短语。那么，怎样的单词或词组才能算作新词语？这在不同的情况下有不同的定义，但至少应当满足以下特点：(1)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[10]；(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词汇”[11]；(3)新词语必须受到普遍的认可，被广泛地应用，在语言词汇中站稳了脚跟，而非昙花一现[12]。

其中第(1)条是新词语的涵义，而第(2)、(3)条是新词语的两个基本特点。“基于WEB的计算机领域新术语的自动检测”算法就是针对新术语/新词语的这两个基本特征设计。其核心思想是，首先通过语料库的比对，找到“自某一时间点”以来在计算机类语料库中新出现的词语，即候选新术语，它们满足新词语的第(2)条特征；然后通过考查候选新术语在时间上的频度曲线，找到其中“被广泛地应用，而非昙花一现”的词语，确定为新术语。

该算法中，这两个核心思想分别是通过“候选新术语发现”和“时间序列分析”两部分完成的。要完成这两部分，还需要三个语料库作为基础。它们分别是：

- 通用语料库：是指由不特定于某一领域的文本组成的语料库，其中的词语为人们所常用。
- 计算机领域背景语料库：进行新术语的检测，需要确定一个时间点，作为考查术语是否为新出现的标准。而计算机领域背景语料库，就是指计算机领域中，该时间点之前出现文本所组成的语料库。所谓“背景”，意指新术语是与该语料库进行比对之后，才被突出出来的。接下来如果不作特殊说明，将简写为“背景语料库”。
- 计算机领域即时新闻语料库：是指在新术语检测时间点之后的新闻文本组成的语料库。该语料库是以时间为单位存放新闻文本的。该语料库是由定时(如每隔3个小时)从英文IT网站获取的新闻文本组成的。接下来如果不做特殊说明，将简写为“即时新闻语料库”。

[†] http://www.rdues.liv.ac.uk/welcome.html。

在以上三个语料库的基础上，算法流程进行如下。图1为该算法流程图。

1) 首先，进行N元词串统计。统计通用语料库和背景语料库中单词和二元词串，并按照其频度进行排序，得到单词表和二元词串表。
2) 其次，进行候选新术语发现。即时新闻所使用的词语，主要来源于三个：通用语料库；背景语料库和第一次出现的词语。因此，这一步骤主要工作就是将即时新闻语料库中新闻文本中，在通用语料库、背景语料库中已经出现的词语去除，所得到的剩下的词串即候选新术语。
3) 最后，进行时间序列分析。对于候选新术语，只考查它们是否具有新词语的一个特点：“在某个时间点之后出现”，但要被认定为新词语，还需要满足另外一个特点：“新词语必须受到普遍的认可，被广泛地应用，在语言词汇中站稳了脚跟，而非昙花一现。”因此需要“时间序列分析”来完成考查候选新术语是否具备这一特点。

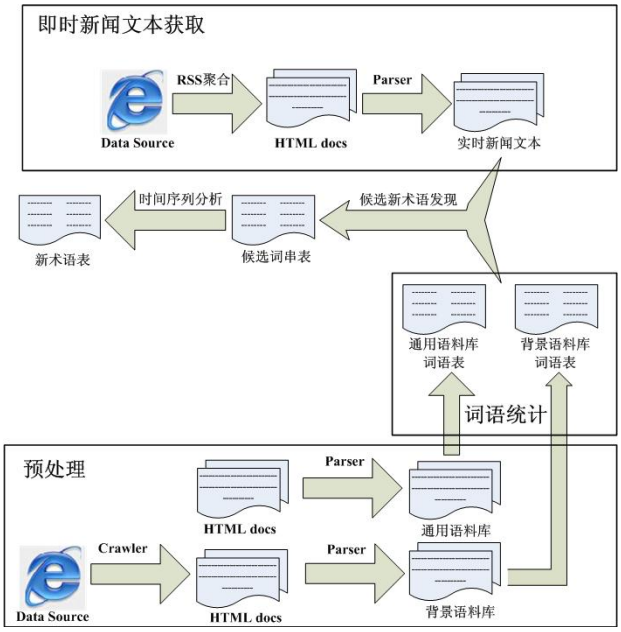

图 1：基于 WEB 的计算机领域新术语自动检测算法流程图

3.1. 预处理

预处理的主要工作是获取通用语料库和背景语料库的原始数据，然后对两部分的原始数据进行解析，获取正文文本，构成通用语料库和背景语料库。通用语料库的原始数据采用英美小说文本《英文世界名著1000部》的English Literature部分。背景语料库的原始数据采用从五大英文IT网站抓取的网页。下面主要介绍如何生成背景语料库。背景语料库的原始数据来源于五个权威英文IT类网站(CNet，ZDNet，Slashdot，PC World，PC Magazine)。最终得到通用语料库数据量为296MB。通过网络爬虫程序，在2006年3月2日到2006年3月7日的时间内，完成对五个IT网站的历史网页的抓取，经过HTML Parser工具，获取背景语料库数据量2.40GB。

3.2. 即时新闻文本获取

如算法流程图所示，算法的“即时新闻文本获取”部分用来生成“即时新闻语料库”。该部分主要包括以下两个步骤，其中第2步如前述第3.1节：(1)定期从五大IT新闻网站发布的新闻RSS源获取新闻网页；(2)对获取的新闻网页，通过HTML Parser提取新闻文本。本研究监控的五

大英文IT网站都提供了丰富的RSS Feeds，涵盖了网站各栏目。这样的优点是：RSS Feeds发布的新闻提供了格式良好的信息，包括标题、链接和发布时间，可以及时增量式地对每日发布的新闻进行获取和存储。

3.3. N元词串统计

这一部分的主要功能是，通过统计的技术，提取通用语料库和背景语料库的特征，在下一流程“候选新术语发现”部分中，用来进行不同语料库之间的比对，提取候选新术语。

在该部分主要统计单词和二元词串的TF和DF。为了避免使统计数据过于庞大，在进行N元词串统计的过程中，引入了类似于Stop Word List的概念——“Frequent Word List”，其中包含从Brown Corpus中统计出来的频度最高的5000个单词[‡]，引入的目的是减小统计过程的时间和统计结果的空间，加快统计的运行速度。最后对通用语料库和背景语料库词语的统计结果如表格3：

表格 1：通用语料库和背景语料库统计表

文件名	数目
通用语料库单词表	290259
背景语料库单词表	531369
通用语料库二元词串表	3845050
背景语料库二元词串	2912505

3.4. 候选新术语发现

“候选新术语发现”部分的主要功能是通过语料库之间的比对，从即时新闻语料库中提取候选新术语，供下一流程“时间序列分析”对其进行分析，最终确定新术语。该流程主要采取了孤岛词串发现算法。孤岛词串发现算法主要功能是：发现孤岛词串，对孤岛词串进行修整，并最终确认输出候选新术语。算法流程如下图所示：

图 2：孤岛词串发现算法示意图

此算法由三个步骤组成，在详细叙述之前，需要事先引入几个词语表和定义几个集合。首先在该部分的“孤岛词串发现算法”中需要使用大量的词语表，它们分别是：(1)通用语料库二

[‡] http://www.edict.com.hk/textanalyser/wordlists.htm

元词串表、背景语料库二元词串表；(2)Frequent Word List（高频词表）；(3)Stop Word List（停词表）：采用 Google 提供的 Stop Word List[§]其中只含有少量的人称代词，助词，冠词等；(4)Empty Word List（虚词表）：该词表来源是 DVL/Verity Stop Word List[**]，并将其中的动词删除；(5)通用语料库单词表、背景语料库单词表；(6)Current IT Term List：该词表包括目前收集到的计算机类新术语。它有两个来源：一个来源是 IT 类网站 PC Magazine 的 IT Encyclopedia[††]，其中收集了超过 2000 个 IT 类术语。另一个来源是该算法发现新术语会自动存放到该术语表中。这个词表主要用来对修整后的孤岛词串，查看其是否已经是之前发现过的（新）术语。以下是根据几个词语表的内容定义的几个集合：

- 设背景语料库二元词串表中频度最高的前 10000 个二元词串集合为 $\mathbb{O}$，通用二元词串表中频度最高的前 10000 个二元词串集合为 $\mathbb{G}$，Frequent Word List 中单词组成的二元词串集合为 $\mathbb{F}$，那么可以对这些集合取并集得 $\mathbb{A}=\mathbb{O}\cup\mathbb{G}\cup\mathbb{F}$。
- 设 Stop Word List 中单词组成的二元词串集合为 $\mathbb{S}$。
- 设 Empty Word List 单词组成的集合为 $\mathbb{E}$。
- 设通用语料库单词表中频度最高的前 5000 个单词集合为 $\mathbb{N}$，背景语料库单词表中频度最高的前 5000 个单词组成的单词集合为 $\mathbb{L}$，Frequent Word List 组成的单词结合为 $\mathbb{R}$，那么可以对这些集合取并集得 $\mathbb{W}=\mathbb{N}\cup\mathbb{L}\cup\mathbb{R}$。
- 设 Current IT Term List 单词组成的集合为 $\mathbb{I}$。

下面详细叙述算法主体三个步骤的具体实现。

第一步：划分二元词串。划分二元词串的过程比较简单。首先将新闻文本转换为小写形式。对其进行划分句子。将每一句子划分为二元词串。对一个句子的单词序列，表示为：$w_1, w_2, w_3, ..., w_n$，可以得到二元词串 $B_1, B_2, ..., B_{n-1}$。很明显，除去句首和句尾的两个单词，其他单词都分别隶属于相邻的两个二元词串。

第二步：初步提取孤岛词串。第二步完成初步提取孤岛词串。这需要用到第 3.3 节“N 元词串统计”部分得到的通用语料库词串表和背景语料库词串表，这一步还要用到 Stop Word List。我们对单词是否在之前出现过，有以下判断：首先对其中任何一个单词 w_i 用以下函数判别：

$$f(w_i)=\begin{cases} Appeared & if\ \ B_{i-1}\in\mathbb{S}\ \text{or}\ B_i\in\mathbb{S} \\ Appeared & if\ \ i=1\ \text{and}\ B_1\in\mathbb{A} \\ Appeared & if\ \ i=n\ \text{and}\ B_{n-1}\in\mathbb{A} \\ Appeared & if\ \ i\in(1,n)\ \text{and}\ B_{i-1},B_i\in\mathbb{A} \\ Unappeared & Others \end{cases}$$

通过该函数可以对该句的每个单词打上一个标签 Appeared/Unappeared，然后将单词序列中标签为 Unappeared 的连续的单词提取出来，这就是孤岛词串。

第三步：修整孤岛词串边界。孤岛词串提取出来之后，需要对其边界进行修整，如图 2 中，Step 2 之后，提取出的“The Affinity Propagation will”中，从直观来讲，“The”、“will”都不应当是新词语的一部分，需要将它们去掉，得到真正的新词语“Affinity Propagation”。这一步骤的详细算法描述如下：

1) 对一个孤岛词串，如果它是一个单词，查看其或者其 Stemming 之后的形式是否属于集合

---

§ http://www.ranks.nl/tools/stopwords.html。

** http://dvl.dtic.mil/stop_list.pdf。除了包括常用助词、代词外，还包括高频动词、副词和形容词等。

†† http://www.pcmag.com/encyclopedia/。

$\mathbb{W}$，如果是，则丢弃，否则进入步骤(3)；如果它是一个二元词串，查看其是否属于集合$\mathbb{A}$，如果是，则丢弃，否则进入步骤(2)；如果它是一个含有超过三个单词的词串，进入步骤(2)；

2) 查看词串的首尾单词是否属于集合$\mathbb{E}$，如果是则将该单词从这个词串中删除，进入步骤(3)；
3) 如果该词串较(1)时有修改，则递归重新进入步骤(1)，对修整后的新词串进行新的一轮修整；否则进入步骤(4)；
4) 查看词串是否属于集合$\mathbb{I}$，即是否出现在 Current IT Term List 中，如果是，则丢弃；否则，输出为新术语候选词串。

3.5. 时间序列分析

该部分对候选词串列表中的词语进行分析，考查候选新术语在一段时间内的频度的变化趋势，最终确定新术语表。时间序列分析的主要功能是对每一个候选新术语，根据其在一段时间内的频度变化，通过一个评价函数，给该候选新术语打分，当分数超过某一个阈值的时候，就最终确定该词语为新术语，提交给用户。

该部分需要提供几个可供调整的参数：(1)开始日期，指分析日期数据的开始点；(2)粒度，指分析的基本单位；(3)分析点数目，指分析的基本单位个数，如以7天为单位进行分析，那么需要确定分析几个7天的数据。

时间序列分析算法如下。首先设定时间序列分析的开始日期为s，粒度为g，分析点数目为n。需要进行以下两个步骤实现该部分功能：第一步，数据聚合。读出日期s的候选新术语，构成候选新术语集合$\mathbb{C}$；对$\mathbb{C}$中每个词语t，查看其在从s开始，$g \times n$天内的频度，得到$g \times n$组频度数据，对该频度数据每g为一组，进行聚合，这里的操作方式是进行简单的求均值。第二步，数据评价。对聚合数据要进行评价，通过所得的分数与阈值比较，判定该候选新术语是否达到新术语的标准。设候选新术语t对应的频度数据为$a_1, a_2, ..., a_n$，经过一些分析，确定评价函数如下：

$$f(a_{i+1}, a_i) = \begin{cases} 1 & \text{if } a_{i+1} > a_i \\ -0.5 & \text{if } a_{i+1} = a_i \\ -1 & \text{if } a_{i+1} < a_i \end{cases}$$

利用上述评价函数，设阈值为δ，评价某候选新术语的频度数据方法如下：(1)对频度数据$a_1, a_2, ..., a_n$的任意$i \in [1, n)$，计算$f(a_{i+1}, a_i)$；(2)通过以下求和公式得到评价函数分数：$S = \sum_{i=1}^{N} f(a_{i+1}, a_i)$，在$N \to n-1$过程中，一旦$S > \delta$即判定该词为新术语。

4. 实验分析

设粒度为1，分析点数目为10，考察2006年5月17日到20日的算法返回新术语。其中发现了若干新术语如“Enterprise Anti-matter”，“Oracle ONE”，“Wearable Tech”，“googlepage”等，表明该算法可以有效地检测到新术语。下面通过手动方式对这四天的算法返回新术语和人工确定新术语的来源进行统计。现在把算法返回新术语的来源分为五个部分：广告、用户ID、新闻正文（包括标题）、用户评论和其它（指导航栏、链接的锚文本等）。

表格 4：算法返回新术语来源分类，第一个数据为个数，第二数据为百分比（%）

日期	广告	ID	新闻正文	评论	其它	总计
2006.06.17	58/46.0	8/6.3	26/20.6	7/5.6	27/21.4	126
2006.06.18	51/66.2	4/5.2	11/14.3	4/5.2	7/9.1	77

2006.06.19	14/40.0	2/5.7	10/28.6	1/2.9	7/20.0	35
2006.06.20	89/50.3	2/1.1	40/22.6	0/0.0	47/26.6	177

表格 5：人工确定新术语来源分类，第一个数据为个数，第二数据为百分比（%）

日期	广告	ID	新闻正文	评论	其它	总计
2006.06.17	6/35.3	0/0.0	8/47.1	0/0.0	3/17.6	17
2006.06.18	9/75.0	0/0.0	1/8.3	1/8.3	1/8.3	12
2006.06.19	1/16.7	0/0.0	4/66.7	0/0.0	1/16.7	6
2006.06.20	5/23.8	0/0.0	6/28.6	0/0.0	9/42.9	21

通过上表可以看出，算法返回的新术语中，来源于广告的占了近半数。主要原因在于，同一支广告，在发布期间，会在多个新闻网页中出现。这些广告虽然可能在每个新闻网页中只出现一次，但会在该天的多个新闻网页中都有出现，这会让该广告中的词串具有较高的TF，因此广告中的候选新术语也在“时间序列分析”中较有可能被判定为新术语。其次，目前新闻网站都会在新闻网页上提供网友评论新闻的功能，因此在每个新闻文本中，还会有大量的用户评论。这就会导致一些比较活跃的网友ID被误认为新术语，

5. 结论和展望

实验表明，该算法已经能够自动检测到相当的新术语。该算法检测新术语的时候，由于采取比较保守的手段，让所有可能的词串加入到该列表，这样可以保证算法的召回率，让用户放心地只需要在列表中甄别新术语即可，无需担心出现的新术语不被该列表收入。但目前该算法实现仍然比较粗糙。未来的工作主要是在保证较高召回率的前提下，不断的提高准确率，即进一步降低用户的查找新术语的工作，包括改进时间序列分析算法等。另外可以看到，算法的准确率较低主要的原因为：第一，大量的广告词在一段时间内反复出现在新闻网页中，会对数据的可靠性造成一定的影响；第二，活跃用户的ID等也是造成算法准确率较低的重要原因。这是今后需要着重分析和解决的。

参考文献

1 汪榕培,陆晓鹃. 英语词汇学教程. 上海: 上海外语教育出版社. 1997

2 张德鑫. “水至清则无鱼”——我的新生词语规范观. 北京大学学报(哲学社会科学版), 2000. 37: 106-119

3 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

4 罗盛芬,孙茂松. 基于字串内部结合紧密度的汉语自动抽词实验研究. 中文信息学报, 2003. 17(3): 9-14

5 周正宇,李宗葛. 一种新的基于统计的词典扩展方法. 中文信息学报, 2001. 15(5): 46-51

6 邹纲, 刘洋, 刘群, et al. 面向 internet 的中文新词语检测. 中文信息学报, 2004. 18: 1-9

7 贾自艳,史忠植. 基于概率统计技术和规则方法的新词发现. 计算机工程, 2004. 30(20): 19-21

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

9 郑家恒, 李鑫,谭红叶. 基于语料库的中文姓名识别方法研究. 中文信息学报, 2000. 14(1): 7-12

10 亢世勇. 《新词语大词典》前言. In: 新词语大词典. 上海: 上海辞书出版社, 2003

11 高永伟. 近 20 年英语国家对新词的研究. 外语与外语教学(大连外国语学院学报), 1998. 114: 8-10

12 刘叔新. 汉语描写词汇. 北京: 商务印书馆. 1990