基于统计的无词典分词方法
Word Extraction without Dictionary Based on Statistics

傅赛香 袁鼎荣 黄柏雄 钟 智
Fu Saixiang Yuan Dingrong Huang Boxiong Zhong Zhi

(中国科学院计算技术研究所智能信息处理开放实验室 北京 100080)
(The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing, 100080)

(广西师范大学计算机科学系 桂林 541004)
(Department of Computer Science, Guangxi Normal University, Guilin, 541004)

摘要 通过分析词的结合模式,提出无词典分词模型,并对该模型进行实验测试。测试结果表明,无词典分词模型能够满足快速分词的要求。

关键词 自动分词 无词典分词 词条过滤 词条统计
中图法分类号 TP391.1 A

Abstract The method for extracting words without dictionary based on statistics is discussed. Three principles on word filtering are proposed by analyzing the combination models of words. The model for word extraction is developed, and meet the requirement of rapid extraction in the experiments.

Key words automated word extraction, word extracting without dictionary, word filtering, word statistics

文档的自动分词一直是中文信息处理技术研究的热点和难点。汉语信息处理系统只要涉及句法、语义(如检索、翻译、文摘、校对等应用),就需要以词为基本单位。例如,汉字的拼音一字转换、自然语言理解、机器翻译、文本分类、汉语文章的自动朗读(即语音合成)、文本校对等中文信息处理系统同样需要分词作为其最基本的模块。因为汉字字符数量多,编码方式复杂,词与词之间却没有分隔符,因此,正确地切分词语,是个很重要的问题。

目前的分词方法归纳起来有3类:第一类是基于语法和规则的分词法。其基本思想就是在分词的同时进行句法、语义分析,利用句法信息和语义信息来进行词性标注,以解决分词歧义现象。因为现有的语法知识、句法规则十分笼统、复杂,基于语法和规则的分词法所能达到的精确度远远还不能令人满意,目前这种分词系统还处在试验阶段。第二类是机械式分词法。机械分词的原理是将文档中的字符串与词典中的词条进行逐一匹配,如果词典中找到某个字符串,

2002-06-08 收稿。

则匹配成功，可以切分，否则不予切分。基于词典的机械分词法，实现简单，实用性强，但机械分词法的最大的缺点就是词典的完备性不能得到保证。据文献[1]统计，用一个含有70 000个词的词典去切分含有15 000个词的语料库，仍然有30%以上的词条没有被分出来，也就是说有4 500个词没有在词典中登录。第三类是基于统计的方法。基于统计的分词法的基本原理是根据字符串在语料库中出现的统计频率来决定其是否构成词。

基于统计的分词方法即是无词典分词方法。本文在分析词的结合模式的基础上，提出3个过滤原则对词条进行过滤，建立无词典分词模型，并对无词典分词模型进行实验测试。

1　无词典分词模型

词是字的组合，相邻的字同时出现的次数越多，就越有可能构成一个词。因此字与字相邻共现的频率或概率能够较好的反映它们成为词的可信度。

1.1　词条模式

如果任意2个或2个以上的汉字组成的连续字符串称为汉字的结合模式，那么，词就可以看成是汉语中字与字的一种结合模式。当然，并不是每一个汉字结合模式都能构成一个词，它必须满足一定语法规则并且具有确定的语义才能称为词。

将词条分为3种模式：统计模式、词法模式、语境模式。

定义1　统计模式定义为一个在文档中出现具有一定的频度的连续的字符串。

比如，从“元搜索引擎”中抽出的“元搜”、“搜索”、“索引”、“引擎”、及“元搜索”、“搜索引”、“索引擎”、及“元搜索引”、“搜索引擎”和“元搜索引擎”共10个字符串，都是具有统计模式的词条。

定义2　词法模式指的是满足一定的语法规则，并具有一定的含义和语义的字符串。

比如，上述的“搜索”、“索引”、“引擎”、“搜索引擎”、“元搜索引擎”共5个字符串就是具有词法模式的词条。

定义3　语境模式指的是，这个词条首先要具有词法模式，并且只能根据上下文的语义和语境才能确定的真实的词或词组。

在上述的词条中，根据上下文，我们只能得到“元搜索引擎”1个是正确的词条，我们称“元搜索引擎”具有语境模式。

从理论上来讲，要取得语法模式、语境模式，非得要有语言学、语法学的参与不可。但根据词条共现的规律，真实词条在文档中共同出现的次数比非真实词条出现的次数要高，也可以利用统计学的知识部分地解决这些问题。根据以上的定义可知，无词典的分词法实际上就是如何统计和过滤的过程。

1.2　词条统计

下面我们给出无词典分词法的基本步骤：(1)对文档进行预处理，包括：编码转换、全半角处理、字符转换，再将停用词、标点符号、英文字母、数学运算符等其它非汉字字符用空格代替。(2)取出所有相连2个或2个以上的汉字组成的字符串，作为本文档的汉字结合模式集，记为Φ，并统计这些字符串的出现频度。去掉频度小于一定阈值的字符串，得到统计模式的词条集合候选集Φ。(3)对候选集Φ中的字符串进行分析，过滤掉不满足其它条件的词条，最后得到文档的识别结果集Ψ。(4)利用最大正向匹配法对识别结果集Ψ中的每个词条进行词频统计。

文档经过预处理以后，一篇文档可以看成是一个字符串，$DOC=S_1S_2\cdots S_n$。其中任何一个

字符 S_i 要么是汉字，要么是空格。不失一般性，不妨假设文档是一个不含空格的字符串 DOC=$S_1S_2\cdots S_n$，其中任意字符 S_i 不为空格。

假定真实词条的最大长度为 Len，则词条统计结合模式的算法如下。

```
Φ=∅(空集),p=1;                    //p 为截取字符串首址
while(p<=len-1){                  //q 为字符串长
  q=1
  while(q<Len-1)&& q<=len-p){
    取词串 strChar=S_p…S_{p+q};
    if(strChar∈Φ)  Frq(strChar)++;
    else(Φ=Φ∪(strChar);Frq(strChar)=1;q++}
  p++;}
```

上述算法在进行字符串查找扫描时，需要进行的匹配字符串的操作次数平均为 $\log_2(i-1)$次，但考虑到文档的长度是有限的，$\log_2(i-1)$应该不大于一个常数，所以算法的时间复杂度为$O(n)$，这完全可以满足大规模中文分词的需要。

1.3 词条过滤

采用无词典分词法，得到的非真实词条是非常多的(约占92.29%)，是真实词条的12倍左右。分析非真实词条的结合模式，可以将其分成几种类型：(1)不相干组合型。比如在文档 D2中的字符串“推出了新的搜索引擎”，除了真实词条外，还有词条“出了”、“了新”、“的搜”等词条，这类词条由于是这些字偶然连在一起而被算法提取出来。这类不相干的词串组合成词频往往很低，绝大多数词频为1，少数词频为2以上。(2)词段组合型。词段组合指的是在得到一个长词条的同时，会得到其很多相应短词条的组合。比如：“电子商务”，在得到“电子商务”这一词条的同时，也会得到“电子商”、“子商”、“子商务”等非真实词条。这类部分组合的字串其词频与其相应的真实词条的词频相同或相近。但是，象“电子”、“商务”两个字串却是真实词条。其词频一般不小于“电子商务”的词频，要区分对待这些词段组合。(3)冗余组合型。冗余组合是由于在取字串时加进了别的字而形成的非真实词条。比如：“搜索引擎”是真实词条，但“搜索引擎是”、“搜索引擎中”却是非真实词条，这类词条的词频虽不是很高，但往往不会太低的。而且，有时这种冗余型组合未必不是真实词条。比如“元搜索引擎”却是真实词条。

对于不相干组合，我们可以设置一个词频阈值，比如词频大于2的字符串才进一步进行筛选。对于词段组合型和冗余组合型的非真实词条，我们采用如下定义和方法：

定义4 词条的支持度指的是词条的词频，即出现在文档中的次数，简记为 sup(w)。

定义5 已知词条 w1的支持度为 sup(w1)，词条 w2的支持度为 sup(w2)，词条 w=w1+w2的支持度为 sup(w)，则词条 w1相对于词条 w 来说，词的置信度为

$$\mathrm{conf}(w1/w)=\frac{\sup(w1)-\sup(w)}{\sup(w1)},$$

同样可知词条 w2相对于词条 w 的置信度。

定理1 取大原则：如果词条 w1相对于词条 w 的置信度小于阈值 α ($\geqslant 0$)，则认为词条 w 是真实词条的可能性比 w1大，从候选集中去掉 w1词条。

比如，w1=“搜索引”，w=“搜索引擎”，sup(w1)=119，sup(w)=119，则 conf(w1/w)=(119−119)/119=0，有 conf(w1/w)小于阈值 $\alpha=0.2$，所以，从候选集中去掉“搜索引”词条，

保留词条“搜索引擎”。

定理2　取小原则：如果词条w1相对于词条w的置信度大于阈值β($\geqslant 0$)，则认为词条w1是真实词条的可能性比w大，从候选集中去掉w词条。

比如，w1=“用户”，w=“给用户”，sup(w1)=46，sup(w)=3，取$\beta=0.8$，则conf(w1/w)=(46−3)/46=0.93$>\beta$。所以，应从候选集中去掉“给用户”词条，保留“用户”词条。

定理3　取中原则：如果词条w1相对于词条w的置信度大于阈值α且小于阈值β，则保留2个词条。

比如，“信息”、“检索”和“信息检索”3个词条的词频分别为137，120，33，其置信度分别为0.759，0.725，置信度大于阈值$\alpha=0.2$且小于阈值$\beta=0.8$，于是可知，这3个词条是真实词条的可能性是很大的。

综上所述，得出无词典分词模型(图1)。

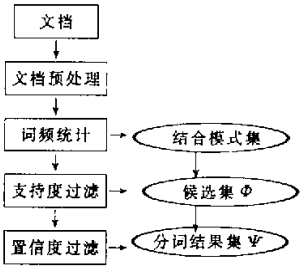

图1　无词典分词模型

2　模型测试

为了测试无词典分词模型的效果，我们进行了小型的测试。

从互联网上收集近40篇文档，文档大小从1K～18K不等。表1、表2给出了部分文档的分词情况。表1、表2中分准率指的是被抽取出来的真实词条(集合W)的数目与识别结果集中词条(集合Ψ)的数目的比值；分全率指的是被抽取出来的真实词条(集合W)的数目与文章中所有真实词条(集合Ω的数目的比值。

表1　文档大小与精确度的关系(取最小词频为3)

文档	大小	识别词条数	错误数	分准率(%)	分全率(%)
D1	1.01	12	1	91.6	9.7
D2	1.51	17	2	88.2	34.2
D3	3.64	96	14	85.4	51.8
D4	8.21	157	20	87.3	27.1
D5	8.88	156	26	83.3	25.3
D6	9.23	185	18	90.3	35.9
D7	14.9	266	34	87.2	31.8
D8	15.3	249	42	83.1	24.7
D9	18.7	337	50	85.1	28.2

表2　文档大小与精确度的关系(取最小词频为4)

文档	大小	识别词条数	错误数	分准率(%)	分全率(%)
D1	1.01	8	0	100	7.1
D2	1.51	8	1	87.5	15.9
D3	3.64	59	5	91.5	34.2
D4	8.21	101	8	92.7	18.4
D5	8.88	110	14	87.3	18.7
D6	9.23	119	11	90.8	14.6
D7	14.9	184	22	88.0	22.2
D8	15.3	181	13	92.8	20.0
D9	18.7	245	19	92.2	22.2

从表1、表2可以看出，无词典分词的分准率与文档的大小关系不大，其分准率都能达到较为满意的效果(80%以上)，但是其分全率却很少有超过50%的。究其原因，根据引理Zipf定律[2]可知：词频为1的词条约占了全部真实词条的一半，而无词典分词法却未能抽取出这样的词条，所以分全率一般不超过50%。对于词频较高的词条，识别效果较好。当所取的最小词频增加时(即考虑中高频词)，可见其分准率也随之提高，但其分全率随之下降。

分词的程序在经过简单的优化之后，分词速度也不错，表3是分词速度的测试。从表3中可知，无词典分词法是能够满足快速分词的要求的。

(下转第264页)

库中抽取满足难度系数等参数要求的试题，在试题多于一道时，随机抽取一道。算法结束后，若fail为1，则组卷失败，否则试题集保存在临时试卷表中。

3 结束语

与传统自动组卷算法不同，本文的算法把知识点的掌握程度值作为组卷的优先级，根据测试知识点掌握程度向量计算测试知识点权值以及确定测试知识点对应的难度系数范围，最后基于知识点循环抽取满足控制参数要求的试题。实验表明，在试题库设计合理、学生模型已建立的前提下，该算法运行稳定，效果良好。

参考文献

1 肖志辉，张祖荫，韩少杰. 智能出卷测试系统的设计与实现. 计算机工程与应用，2000，10：84～99.

2 林雪明，张均良，蒋伟钢. 基于知识点的试题组卷算法的建立. 微机发展，2001，2：77～79.

3 乐毓俊，刘占平，刘光然. 智能教学系统集成开发环境及其认知型学生模型的研究与实现. 宁夏大学学报（自然科学版），1996，16(4)：20～28.

4 苏德富，钟 诚. 计算机算法设计与分析. 北京：电子工业出版社，2001.

（责任编辑：蒋汉明）

（上接第255页）

表3 分词速度测试

文档	文档大小	识别词条数	分词时间	速度（词/秒）
D3	3.64	59	0.051	1157
D4	8.21	101	0.099	1020
D5	8.88	110	0.103	1068
D6	9.23	119	0.112	1062
D7	14.9	184	0.171	1076
D8	15.3	181	0.169	1071
D9	18.7	245	0.214	1144
总计	78.86	999	0.919	平均1087

3 结束语

当然，无词典分词法也有一定的局限性，会经常抽出一些共现频度高、但并不是词的常用字符串，如“这一”、“之一”以及“提供了”等等。在实际应用的统计分词系统中都要使用一部基本的分词词典（常用词词典）进行串匹配分词，即将字符串的词频统计和字符串匹配结合起来，既发挥匹配分词切分速度快、效率高的特点，又利用了无词典分词结合上下文识别生词、自动消除歧义的优点。

参考文献

1 Chien Lee-Feng. PAT-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval. Information Processing and Management，1999，35：501～521.

2 ZIPF H P. Human Behaviour and the Principle of Least Effort. Addison-wesley，Cambridge，Massachusetts，1949.

（责任编辑：邓大玉）