# 最大匹配算法研究

万 莛

（武汉理工大学 计算机科学与技术学院，湖北 武汉 430063）

**摘 要**：最大匹配算法是中文分词中最常用的方法，但其有着过分依赖于词典的弊端。对最大匹配算法进行了深入探讨与研究，使用 n-gram 技术更新词典解决其弊端，从而提高分词效果。最后通过双向匹配算法与 n-gram 相结合的实验验证了该方案的可行性，并对该方案进行了总结。

**关键词**：中文分词；最大匹配；n-gram；词频；双向匹配

**中图分类号**：TP391.1 **文献标识码**：A **文章编号**：1674-7720(2012)08-0062-02

## The research of maximun match algorithm

Wan Ting

(College of Computer Science and Technology, Wuhan University of Technology, Wuhan 430063, China)

**Abstract**: As the most common algorithm in Chinese segment, maximun match algorithm also has a problem that excessive depend on the dictionary . After the intensive study of maximun match algorithm ,this paper presents a way to update the dictionary by using the technology of n-gram to solve its problem. Thus improve the quality of maximum match algorithm. Finally do a test that use bilateral match and n-gram technology together to prove the way is feasible and the way is summarized.

**Key words**: Chinese segment; maximum match; n-gram; word frequency; bilateral match bilateral match

作为计算机信息处理中最基础、最关键的技术，中文分词一直是人们研究的热点。中文分词就是将连续的汉字序列按照一定的规律分割成一个个单独的词的过程[1]。在英文句子中，单词之间是以空格作为自然分界符的，所以英文分词比较简单；而中文以字为基本单位，将一序列字串联在一起形成句子，从而表达意思，中文的句和段能通过明显的分界符来划分，但是词没有一个形式上的分界符，所以中文分词比英文分词相对困难许多。中文分词方法总结起来大致可分为三大类：基于词典直接匹配的分词方法、基于规则和理解的分词方法和基于统计模型的分词方法[2]。本文主要讨论基于词典匹配算法中的最大匹配算法，针对其过分依赖词典这一弊端进行了探讨并提出了对策。

## 1 最大匹配算法

最大匹配算法是最常用也是最基本的字符串匹配算法之一。它能够保证切分出来的词长度最大，同时易于实现[3]。最大匹配算法包括正向最大匹配算法、逆向最大匹配算法和双向最大匹配算法。

### 1.1 正向最大匹配算法

正向最大匹配算法流程[4]如图 1 所示。

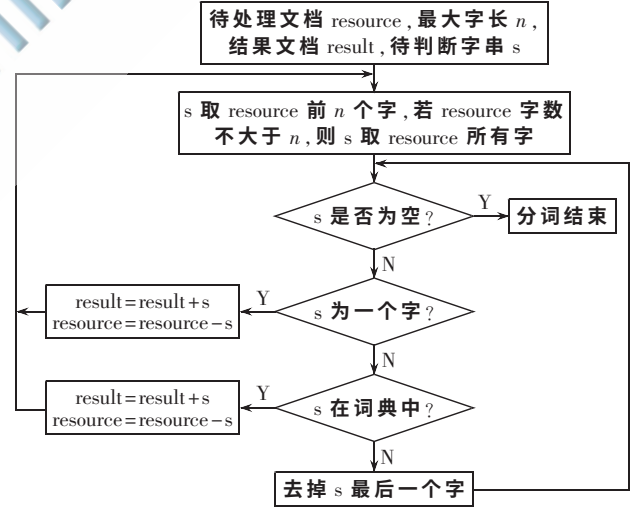


图 1 正向最大匹配算法流程图

以“中华人民共和国简称中国”为例，设定取词长度 $n$ 为 8，待匹配字符串为 s，按照上述步骤处理过程为：

(1)s 为“中华人民共和国简”，查找词典进行匹配操作，发现没有该词；

(2)s 去掉最后一个字，变为“中华人民共和国”，查

找词典进行匹配操作，发现该词，将该词存入结果文档中；

(3)更新 s，发现剩余的字“简称中国”长度不足 8，所以 s 为“简称中国”，查找词典进行匹配操作，发现没有该词；

(4)s 去掉最后一个字，变为“简称中”，查找词典进行匹配操作，发现没有该词；

(5)s 去掉最后一个字，变为“简称”，查找词典进行匹配操作，发现该词，将其存入结果文档中；

(6)更新 s，发现剩余的字“中国”长度不足 8，所以 s 为“中国”，查找词典进行匹配操作，发现该词，将其存入结果文档中；

(7)更新 s，发现 s 为空，至此分词操作结束。

分词结果为“中华人民共和国/简称/中国”。

### 1.2 逆向最大匹配算法

逆向最大匹配算法与正向最大匹配算法流程相似[5]，只是取词操作与待匹配字串更新操作不同。逆向最大匹配算法从文档末尾开始进行取词，匹配不成功删除的是待匹配字符串的第一个字而不是最后一个。

### 1.3 双向最大匹配算法

将正向最大匹配算法与逆向最大匹配算法相结合所产生的算法即是双向最大匹配算法，它能够选取正向最大匹配算法和逆向最大匹配算法中分词效果较好的一方，以提高分词效果。

### 1.4 最大匹配算法的问题

最大匹配算法存在以下问题：(1)待匹配字符串最大长度的设定困难，过长易造成效率低，过短则造成分词不精确；(2)对词典依赖程度过大，分词效果取决于词典。

## 2 n-gram 技术

随着时间推移，肯定有大量新词产生。为了丰富词典，本实验采用 n-gram 技术扩充词典。n-gram 就是对一个字序列进行分割，分割产生的字符串是该字序列的子串[6]。例如：对“中华人民共和国简称中国”进行 n-gram 2 元切分，得到 2 元组：中华|华人|人民|民共|共和|和国|国简|简称|称中|中国。

本实验中，n-gram 扩充词典步骤[7]如下：

(1)选择语料库，本次实验选择 2010 年 10 月 14 日~18 日参考消息作为预料库；

(2)对语料库进行预处理，将数字、标点、字母等全部删掉，只剩下汉字；

(3)进行切分并统计词频，本次实验最大词长为 8，所以切分为 2 元组至 8 元组，词频统计如表 1 所示。

表 1 词频数所占比例

| 词频 | 1 | 2 | 3 | 4 | 5 | 大于 5 |
|---|---|---|---|---|---|---|
| 比例/% | 85.693 | 11.214 | 1.244 | 0.806 | 0.254 | 0.789 |

(4)选取候选词。如图 2 所示，根据观察，当设置使用词频大于 5 的词作为候选词时，可靠性较高。所以本次实验设定候选词的词频数大于 5。

图 2 部分词频数大于 5 的词

(5)使用候选词对词典进行更新。

## 3 实验

为了提高分词效果，本次实验采用双向匹配算法同时使用 n-gram 技术来负责词典的更新操作。

由于逆向最大匹配算法比正向最大匹配算法有更高的精度[8]，所以本次实验中双向匹配算法的选择策略是：当正向最大匹配分词所分出的词数小于逆向最大匹配算法所分出的词数时，分词结果采用正向最大匹配所得结果；否则使用逆向最大匹配算法所得结果。

本次实验代码采用 java 编写，分词算法中使用的方法主要有：(1)public StringBuffer result (String s,Set<String> dic)。用一个 StringBuffer 来存储结果，并返回。参数 s 表示从待分词文档中读到的行字符串，dic 表示词典。(2)public void segment (String max,String s, Set<String> dic)。参数 max 表示待匹配字符串。(3)public void n_gram()。n_gram 的实现，主要使用 HashMap<String, Integer>，其中 String 用来存储词，Integer 存储词频。

在未进行词典更新操作之前，对“胡锦涛提出了科学发展观”进行分词操作得到的结果是“胡锦涛/提出/了/科学发展/观”；进行词典更新之后，“科学发展观”成为单独一词，结果为“胡锦涛/提出/了/科学发展观”，说明使用 n-gram 对词典进行更新的确能起到提高分词效果的作用。

本文首先对最大匹配算法进行了详细的阐述，继而提出最大匹配算法的不足之处，即对词典依赖程度过大，词典的好坏直接决定了分词的质量。为解决该问题，提出使用 n-gram 技术来进行词典的自我更新，提高词典质量，从而提高最大匹配算法分词效果。通过实验验证了该方法的可行性。但是仍有不足之处：第一是对用来更新词典的语料库要求较高，语料库必须具有代表性，能包含当前社会所使用的主流词语；其次语料库必须足够大才能得到更好的效果；再次更新词典对计算机性能消耗较大，必须选择合理的时间进行更新操作。

参考文献

[1] 周宏宇，张政.中文分词技术综述[J].安阳师范学院学报，2010(2)：54-56.

[2] 刘春辉.基于优化最大匹配的中文分词方法研究[D].秦皇岛：燕山大学，2009.

[3] 林浩，韩冰，杨乐华.一种基于改进最大匹配快速中文分

词算法[J].科技创新导报,2009(9):248.
[4] 赵源.基于最大匹配的中文分词改进算法研究[J].科技信息,2010(35):487,496.
[5] 王瑞雷,栾静,潘晓花,等.一种改进的中文分词正向最大匹配算法[J].计算机应用与软件,2011,28(3):195-197.
[6] 吴胜远.一种汉语分词方法[J].计算机研究与发展,1996,33(4):306-311.
[7] 李文,洪亲,滕忠坚,等.基于n-gram的字符串分割技术的算法实现[J].计算机与现代化,2010(9):85-87.
[8] 张磊,张代远.中文分词算法解析[J].电脑知识与技术,2009,5(1):192-193.

(收稿日期:2011-12-19)

作者简介:
万莛,男,1988年生,硕士研究生,主要研究方向:软件工程方法与技术。