

数据库、信息处理

## 回溯正向匹配中文分词算法

张劲松, 袁 健

上海理工大学 计算机与电气工程学院, 上海 200093

收稿日期 2008-5-15 修回日期 2008-8-14 网络版发布日期 接受日期

**摘要** 在最大匹配法(The Maximum Matching Method)的基础上,提出了一种回溯正向匹配(The Backtracking Matching Method)的中文分词方法。该方法首先对待切文本进行预处理,将文本分割成长度较短的细粒度文本;利用正向匹配、回溯匹配和尾词匹配来有效发现歧义字段;利用长词优先兼顾2词簇的方式对交集体歧义字段进行切分。最后对该算法进行的定性分析说明了该方法的先进性,从随机抽取大量语料实验结果上证明了该方法的有效性。

**关键词** [中文分词](#) [最大匹配法](#) [回溯匹配](#) [交集体歧义](#)

分类号

## Backtracking matching Chinese segmentation method

ZHANG Jin-song, YUAN Jian

College of Computing & Electrical Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

### Abstract

Chinese word segmentation methods of the backtracking matching method is proposed based on the maximum matching method. The text is pretreatment by the BM in the first, then, the text is cut into shorter lengths granular text; Ambiguity field is found effective by forward matching method, backtracking matching method and last words matching Method; Crossing ambiguity field is cut by Long term priorities and 2-words rules. In the last, the algorithm of the qualitative analysis shows that the method of the advanced nature. The large number of randomly selected language materials being tested and results show that the method is effective.

**Key words** [Chinese word segmentation](#) [Maximum Matching \(MM\) methods](#) [backtracking matching](#) [crossing ambiguity](#)

DOI: 10.3778/j.issn.1002-8331.2009.22.043

通讯作者 张劲松 [sturdys@163.com](mailto:sturdys@163.com)

### 扩展功能

#### 本文信息

- ▶ [Supporting info](#)
- ▶ [PDF\(372KB\)](#)
- ▶ [\[HTML全文\]\(0KB\)](#)
- ▶ [参考文献](#)

#### 服务与反馈

- ▶ [把本文推荐给朋友](#)
- ▶ [加入我的书架](#)
- ▶ [加入引用管理器](#)
- ▶ [复制索引](#)
- ▶ [Email Alert](#)
- ▶ [文章反馈](#)
- ▶ [浏览反馈信息](#)

#### 相关信息

- ▶ [本刊中 包含“中文分词”的相关文章](#)
- ▶ [本文作者相关文章](#)
- [张劲松](#)
- [袁 健](#)