计算机工程与应用 2008 44 (26): 135-137 ISSN: 1002-8331 CN: 11-2127/TP

数据库、信号与信息处理

汉语分词索引字数与分词效率的对比研究

何利益[1,2], 郭 罡[2], 郭建彬[2]
1.中国科学技术大学 电子工程与信息科学系，合肥 230027
2.中国人民解放军 96151部队，安徽 黄山 245041

收稿日期 2007-11-5 修回日期 2008-1-21 网络版发布日期 2008-9-8 接受日期

摘要 针对汉语分词词典中双字哈希索引机制未能充分利用索引分词，而分词效率又明显优于首字哈希索引机制的问题，在充分分析汉语构词特点的基础上，提出了基于三字哈希索引的分词词典机制，并通过将字串的三态标记与下一索引指针的乘积作为哈希值的链地址法，简化了词典结构，节省了内存空间。理论分析和真实语料仿真均证明了三字哈希索引机制与不同字数的其他索引机制相比，具有更好的分词效率。

关键词 计算机应用 中文分词 词典机制 三字哈希索引
分类号

Contrast study on Chinese word segmentation efficiency with different index degree

HE Li-yi[1,2],GUO Gang[2],GUO Jian-bin[2]

1.Department of Electronic Engineering and Information Science，University of Science and Technology of China，Hefei 230027，China
2.96151 Unit of PLA，Huangshan，Anhui 245041，China

Abstract
According to the Chinese dictionary word segmentation efficiency that based on the Double-Character-Hash-Index（DCHI） mechanism exceeds clearly based on the First-Character-Hash-Index（FCHI） mechanism，this paper lucubrates to the Chinese word-building characteristic and provides a new segmentation dictionary mechanism named Three-Character-Hash-Indexing（TCHI） mechanism，which exploits character coding index sufficiently.The results show that the TCHI dictionary mechanism can improve speed and achieve more efficiency than FCHI，DCHI and four；character-hash-index in Chinese dictionary word segmentation mechanism.

Key words computer application Chinese word segmentation dictionary mechanism Three Character Hash Index（TCHI）

DOI: 10.3778/j.issn.1002-8331.2008.26.041

通讯作者 何利益