计算机工程与应用 2007 43 (3期): 179 ISSN: 1002-8331 CN: 11-2127/TP

数据库与信息处理

中文垃圾邮件过滤系统中的实时分词算法设计

申庆永 张建忠 何云 杨洁

南开大学计算机系 南开大学计算机系网络实验室 南开大学计算机科学与技术系 长沙交通学院计算机工程系

收稿日期 2006-2-23 修回日期 网络版发布日期 2007-1-17 接受日期

摘要 在基于内容的中文反垃圾邮件技术中，中文分词是必不可少的一个环节。面对大规模的邮件训练样本和大负载的邮件服务器，中文分词算法的时间效率成为中文垃圾邮件过滤技术中的一个瓶颈。对此，本文提出一种应用在中文垃圾邮件过滤系统中的实时分词算法。该算法采用一种TRIE树型结构作为词典载体并基于最大匹配的原则，同时，在实时分类阶段结合hash表进行特征查询，极大地提高了系统的时间效率。

关键词 中文分词 垃圾邮件 TRIE树

分类号

An Algorithm of Chinese Word Segmentation In Anti-spam System

Abstract

Chinese word segmentation is an absolutely necessary step in the Chinese anti-spam technologies based on mail content. The efficiency of word segmentation algorithm is becoming a bottleneck when it is used in the training of abundant mail samples or on the high load mail server. A real time algorithm is proposed here, which uses a TRIE structure as the carrier of dictionary. Based on the Maximum Matching (MM) principle and combined with the hash table of word attributes, this algorithm improves the efficiency of the anti-spam system observably.

Key words Chinese word segmentation Spam TRIE tree

DOI:

通讯作者 申庆永 sqy@mail.nankai.edu.cn