计算机工程与应用 2010 46 (4): 129-131 ISSN: 1002-8331 CN: 11-2127/TP

数据库、信号与信息处理

章回小说的有意义串发现算法

李海涛，马振华，沈文华

台州职业技术学院 计算机工程系，浙江 台州 318000

收稿日期 2008-8-27 修回日期 2008-11-17 网络版发布日期 2010-2-2 接受日期

摘要 已有有意义串发现算法对于大规模语料中频繁出现的有意义串发现效果较好，而对于语料规模小，或者出现频次较低的有意义串识别效果不够理想。根据章回小说有意义串出现的特点，提出有意义串的局部性原理，并给出了字符串局部性的有效度量方式。将字符串的局部性和语用独立性结合起来，使用局部性和独立性共同描述字符串为有意义串的可能性。实验结果表明：该方法对于章回小说有意义串发现的准确率高于已有方法，同时能够更有效地发现较多的低频有意义串。

关键词 有意义串 章回小说 局部性度量 局部性约束 低频串

分类号 TP301.6

Meaningful string discovery algorithm for chapter-novel corpora

LI Hai-tao，MA Zhen-hua，SHEN Wen-hua

Department of Computer Engineering，Taizhou Vocational & Technical College，Taizhou，Zhejiang 318000，China

Abstract

Available meaningful string discovery algorithms are geared to mining frequent meaningful strings of large-scale corpus.As for small corpus，or less-frequent meaningful strings，their performance is poor.According to the distribution pattern of meaningful strings in chapter-novels，the theory of locality is presented，as well as an effective locality measuring method.Locality and independency are combined to describe the probability of a string to be meaningful.Experiments indicate that the method out-performances all available algorithms.At the same time，the method is able to discover less-frequent meaningful strings effectively.

Key words meaningful string chapter-novel locality measure locality constrain low-frequent string

DOI: 10.3778/j.issn.1002-8331.2010.04.041

通讯作者 李海涛 lht826@gmail.com