计算机工程与应用 2010 46 (7): 148-152 ISSN: 1002-8331 CN: 11-2127/TP

数据库、信号与信息处理

一种用于抄袭识别的文档距离度量

胡明晓[1]，DING Leon X[2]

1.温州大学 计算机科学与工程学院，浙江 温州 325035
2.IBM多伦多实验室，多伦多，ON，L6G 1C7，加拿大

收稿日期 2009-2-26 修回日期 2009-4-8 网络版发布日期 2010-3-2 接受日期

摘要 广义编辑距离的计算是一个NP-完全问题，在充分考虑了文档抄袭行为的特点之后提出一种基于广义编辑距离的单向的低计算复杂性的文档距离度量方法。首先，计算第一文档的各段落在第二文档全文中的近似串匹配距离之和，同时确定各段落在第二文档中的近似匹配子串（即原象串），然后根据这些原象串得到回退数和前跳数，最后将三者求和作为文档距离。该文档距离是一种广义编辑距离的近似值，能够在O（n2）时间内计算，并能充分反映抄袭方向。针对人工文档和实际文档的两组实验表明该距离具有较低的漏检率、误检率。

关键词 文档距离 广义编辑距离 近似串匹配 抄袭识别 电子文档管理

分类号 TP311

Document distance metric used in plagiarism detection

HU Ming-xiao[1]，Leon X.Ding[2]

1.College of Computer Science and Engineering，Wenzhou University，Wenzhou，Zhejiang 325035，China
2.IBM Toronto Laboratory，Toronto，ON，L6G 1C7，Canada

Abstract

The algorithm for generlized edit distance is NP-complete.A one-direction，low complexity document distance metric based on generalized edit distance is proposed after probing special patterns of document plagiarism.Firstly，compute the sum of approximate string matching distances of each paragraph of the first document to the full text of the second document，and determine the best matching substrings in the second document，which is called original map substring，for each paragraph.Then collect returning number and skipping number according to these original map substrings.Finally，sum up the total approximate matching distances，returning number and skipping number to arrive document distance.This document distance metric is an approximation of generalized edit distance，and it can be calculated in O（n2） time and can detect plagiarizing direction.Applications of this new metric on manually created and real-life documents indicate that it has low missing rate and false-alarm rate.

Key words document distance generalized edit distance approximate string matching plagiarism detection electronic document management

DOI: 10.3778/j.issn.1002-8331.2010.07.045

通讯作者 胡明晓 jsj_hmx@wzu.edu.cn