计算机工程与应用 2008 44 (5): 180-183 ISSN: 1002-8331 CN: 11-2127/TP

数据库与信息处理

n-Gram/2L索引结构的存储与时间优化算法

刘凤晨[1], 刘庆文[2], 胡 玥[2], 黄 河[1]
1.北京航空航天大学 软件学院，北京 100083
2.北京科技大学 计算机科学系，北京 100083

收稿日期 2007-6-5 修回日期 2007-8-13 网络版发布日期 2008-1-31 接受日期

摘要 对分词检索算法n-Gram/2L的索引结构作了改进，在第二级倒排表中加入对文章标识的索引，提出一种基于Zigzag的分词检索算法*n*-Gram/2LZ（*n*-Gram/2L on Zigzag join）。在对数据量较大的文章进行检索和索引时，该算法在保留原有算法特性的基础上进一步减少了索引冗余，降低了索引的存储量，同时对查询算法的优化降低了查询时的系统开销，并且减少索引中记录访问次数，提高了查询效率。

关键词 算法 索引 *n*-gram 倒排表
分类号

Space and time optimized algorithm of *n*-Gram/2L index structure

LIU Feng-chen[1],LIU Qing-wen[2],HU Yue[2],HUANG He[1]

1.College of Software，Beijing University of Aeronautics and Astronautics，Beijing 100083，China
2.Department of Computer Science，Beijing University of Science and Technology，Beijing 100083，China

Abstract
This paper presents an improved algorithm of n-Gram/2L index for text retrieval by adding document identifier index into the secondary level inverted index，and proposes a retrieval algorithm：*n*-Gram/2LZ（*n*-Gram/2L on Zigzag join） based on Zigzag join.This algorithm retains the advantage of former *n*-Gram/2L algorithm and reduces redundancy and storage of the document index，while retrieving and indexing large data.And the optimization of the query algorithm decreases the system overhead when processing query as well as enhances query efficiency by reducing reading the same record repeatedly.

Key words algorithms indexing *n*-gram inverted index

DOI:

通讯作者 刘凤晨 aric101@hotmail.com