计算机应用 2013, 33(08) 2208-2211 DOI: ISSN: 1001-9081 CN: 51-1307/TP

数据库技术

海量数据的相似重复记录检测算法

周典瑞,周莲英

江苏大学 计算机科学与通信工程学院，江苏 镇江 212013

摘要： 针对海量数据下相似重复记录检测算法的低查准率和低效率问题，采用综合加权法和基于字符串长度过滤法对数据集进行相似重复检测。综合加权法通过结合用户经验和数理统计法计算各属性的权重。基于字符串长度过滤法在相似检测过程中利用字符串间的长度差异提前结束编辑距离算法的计算，减少待匹配的记录数。实验结果表明，通过综合加权法计算的权重向量更加全面、准确反映出各属性的重要性，基于字符串的长度过滤法减少了记录间的比对时间,能够有效地解决海量数据的相似重复记录检测问题。

关键词： 海量数据 相似重复记录 综合加权法 编辑距离

Algorithm for detecting approximate duplicate records in massive data

ZHOU Dianrui,ZHOU Lianying

School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang Jiangsu 212013, China

Abstract: For the problem of low precision and low time efficiency of approximate duplicate records detection algorithm in massive data, integrated weighted method and filtration method based on the length of strings were adopted to do the approximate duplicate records detection in dataset. Integrated weighted method integrated user experience and mathematical statistics to calculate the weight of each attribute to make weight calculation more scientific. The filtration method based on the length of strings made use of the length difference between strings to terminate the edit distance algorithm earlier which reduced the number of the records to be matched during the detection process. The experimental results show that the weight vector calculated by the integrated weighted method makes the importance of each field more comprehensive and accurate. The filtration method based on the length of strings reduces the comparison time among records and effectively solves the problem of the detection of approximate duplicate records under massive data.

Keywords: massive data approximate duplicate record integrated weighted method edit distance

收稿日期 2013-02-25 修回日期 2013-04-06 网络版发布日期 2013-09-11

DOI:

基金项目:

江苏省科技支撑项目

通讯作者: 周典瑞

作者简介: 周典瑞(1987-)，男，山东泰安人，硕士研究生，主要研究方向：数据清洗;
周莲英(1964-)，女，江苏泰州人，教授，博士，主要研究方向：计算机网络性能分析、信息安全、电子商务、网络信息系统。
作者Email: zdianrui@126.com

参考文献：

［1］MONGE A E, ELKAN C P. The field matching problem: algorithms and applications ［C］// Proceedings of the 2nd Conference on Knowledge Discovery and Data Mining. Cambridge: AAAI, 1996:267-270.

［2］MINTON S N, NANJO C, KNOBLOCK C A, et al. A heterogeneous field matching method for record linkage ［C］// Proceeding of the 5th IEEE International Conference on Data Mining. Piscataway: IEEE, 2005: 314-321.

［3］HERNANDEZ M, STOLFO S. The merge/purge problem for large databases ［C］// Proceedings of

the 1995 ACM SIGMOD International Conference on Management of Data. New York: ACM, 1995:127-138.

［4］BLENK O M, MOONEY R. Adaptive name matching in information integration ［J］. IEEE Intelligent Systems,2003,18(5):16-23.

［5］邱越峰,田增平,季文赟，等.一种高效的检测相似重复记录的方法［J］.计算机学报,2001,24(1):69-77.

［6］鲁均云,李星毅,施化吉，等.基于内码序值聚类的相似重复记录检测方法［J］.计算机应用研究,2010,27(3),874-878.

［7］孟祥逢,鲁汉榕,郭玲，等.基于遗传神经网络的相似重复记录检测方法研究［J］.计算机工程与设计,2010,31(7):1550-1553.

［8］李星毅,包从剑,施化吉.数据仓库中的相似重复记录检测方法［J］.电子科技大学学报,2007,36(6):1273-1277.

［9］MONGE A E, ELKAN C. An efficient domain-independent algorithm for detecting approximately duplicate database records ［C］// Proceedings of the SIGMOD 1997 Workshop on Research Issues on Data Mining and Knowledge Discovery. Cambridge: AAAI, 1997:23-29.

［10］张永,迟忠先.位置编码在数据仓库ETL中的应用［J］.计算机工程,2007,33(1):50-52.

本刊中的类似文章

1．刘雪琼 武刚 邓厚平.Web信息整合中的数据去重方法[J]. 计算机应用, 2013,33(09): 2493-2496

2．刘丽霞 张志强.基于Trie树的相似字符串查找算法[J]. 计算机应用, 2013,33(08): 2375-2378

3．黄国林 郭丹 胡学钢.基于通配符和长度约束的近似模式匹配算法[J]. 计算机应用, 2013,33(03): 800-805

4．蒋新华 廖律超 邹复民.基于浮动车移动轨迹的新增道路自动发现算法[J]. 计算机应用, 2013,33(02): 579-582

5．徐翔 邹复民 廖律超 朱铨.基于GemFire的海量数据计算性能实验分析[J]. 计算机应用, 2013,33(01): 226-229

6．贾楠 付晓东 黄袁 刘晓燕 代志华.基于树编辑距离的工作流距离度量方法[J]. 计算机应用, 2012,32(12): 3529-3533

7．黄亮 赵泽茂 梁兴开.基于编辑距离的Web数据挖掘[J]. 计算机应用, 2012,32(06): 1662-1665

8．夏秀峰 赵龙.基于三层存储模型的RFID数据压缩存储方法[J]. 计算机应用, 2012,32(03): 625-628

9．黄海峰 张珂珩 张鸿 季学纯 陈鹏.电力系统动态信息数据库关键技术[J]. 计算机应用, 2011,31(06): 1681-1684

10．张玉芳 李川 熊忠阳.改进的本体匹配算法[J]. 计算机应用, 2011,31(04): 1067-1069

11．李睿 曾俊瑀 周四望.基于局部标签树匹配的改进网页聚类算法[J]. 计算机应用, 2010,30(3): 818-820

12．王灿 秦志光 冯朝胜 彭静.面向重复数据消除的备份数据加密方法[J]. 计算机应用, 2010,30(07): 1763-1766

13．赵作鹏 尹志民 王潜平 许新征 江海峰.一种改进的编辑距离算法及其在数据处理中的应用[J]. 计算机应用, 2009,29(2): 424-426

14．王春 马纯永 陈戈.基于GPGPU的海量山地地形数据的实时绘制算法[J]. 计算机应用, 2009,29(08): 2105-2108

15．李慧云 殷人昆 冉望.基于海量数据的集群服务管理模型[J]. 计算机应用, 2008,28(5): 1316-1318

16．卫婷 吴渝 李银国.一种可伸缩的粒计算知识获取方法[J]. 计算机应用, 2007,27(9): 2281-2283

17．张建锦 吴渝 刘小霞.一种改进的密度偏差抽样算法[J]. 计算机应用, 2007,27(7): 1695-1698

18．周文 徐国梁.翻译记忆中语句相似度计算方法的研究[J]. 计算机应用, 2007,27(5): 1210-1213

19．张永; 迟忠先; 闫德勤.数据仓库ETL中相似重复记录的检测方法及应用[J]. 计算机应用, 2006,26(4): 880-882

20．杨长辉 岳友友 .一种基于编辑距离的XML查询方案[J]. 计算机应用, 2006,26(12): 2991-2993

21．程国达，苏杭丽.一种检测汉语相似重复记录的有效方法[J]. 计算机应用, 2005,25(06): 1362-1365

22．程国达，邹亚会，朱静.一种自适应信息集成方法[J]. 计算机应用, 2005,25(03): 666-669