P.O.Box 8718, Beijing 100080, China	Journal of Software, Nov. 2006,17(11):2289-2301
E-mail: jos@iscas.ac.cn	ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP
http://www.jos.org.cn	Copyright © 2006 by *Journal of Software*

基于数据网格的书法字k近邻查询

庄 毅, 庄越挺, 吴 飞

庄 毅, 庄越挺, 吴 飞
(浙江大学 计算机科学与技术学院,浙江 杭州 310027)
作者简介: 庄毅(1978－),男,浙江杭州人,博士生,主要研究领域为高维数据查询,网格计算和多媒体检索.庄越挺(1965－),男,博士,教授,博士生导师,CCF高级会员,主要研究领域为多媒体检索,数字图书馆,视频动画.吴飞(1973－),男,博士,副教授,CCF高级会员,主要研究领域为多媒体检索,机器学习.
联系人: 庄 毅 Phn: +86-571-87951853, E-mail: zhuangyi@cs.zju.edu.cn, http://www.zju.edu.cn

Received 2006-06-10; Accepted 2006-08-25

Abstract
In this paper, a novel k-Nearest Neighbor (k-NN) query over the Chinese calligraphic character databases based on Data Grid is proposed. First when user in the query node submits a query character and k, the character filtering algorithm is performed using the hybrid distance metric (HDM) index. Then the candidate characters are transferred to the executing nodes in a package mode. Furthermore, the refinement process of the candidate characters is conducted in parallelism to get the answer set. Finally, the answer set is transferred to the query node. If the number of answer set is less than k, then the query procedure is re-performed by increasing the query radius until the k nearest neighbor characters are obtained. The analysis and experimental results show that the performance of the algorithm is good in minimizing the response time by decreasing network transfer cost and increasing parallelism of I/O and CPU.

Zhuang Y, Zhuang YT, Wu F. Answering k-NN query of Chinese calligraphic character based on data grid. *Journal of Software*, 2006,17(11):2289-2301.
DOI: 10.1360/jos172289
http://www.jos.org.cn/1000-9825/17/2289.htm

摘要
提出一种在数据网格环境下的书法字k近邻查询方法.当用户在查询结点提交一个查询书法字和k时,首先以一个较小的查询半径,在数据结点进行基于混合距离尺度的书法字过滤,然后将过滤后的候选书法字以"打包"传输的方式发送到执行结点,在执行结点并行地对这些候选书法字进行距离(求精)运算,最终将结果书法字返回到查询结点.当返回的书法字个数小于k时,扩大半径值,继续循环,直到得到k个最近邻书法字为止.理论分析和实验表明,该方法在减少网络通信开销、增加I/O和CPU并行、降低响应时间方面具有较好的性能.

基金项目：Supported by the National Natural Science Foundation of China under Grant No.60533090 (国家自然科学基金); the National Science Fund of China for Distinguished Young Scholar under Grant No.60525108 (国家杰出青年基金); the China US Million Book Digital Library Project (高等学校中英文图书数字化国际合作计划)

References:

[1] Palmondon R, Srihari SN. On-Line and off-line handwriting recognition: A comprehensive survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000,22(1):63-84.

[2] Rath TM, Kane S, Lehman A, Partridge E, Manmatha R. Indexing for a digital library of George Washington's manuscripts: A study of word matching techniques. Technical Report, MM-36, Boston: University of Massachusetts, 2002.

[3] Yosef IB, Kedem K, Dinstein I, Beit-Arie M, Engel E. Classification of hebrew calligraphic handwriting styles: Preliminary results. In: Proc. of the 1st Int'l Workshop on Document Image Analysis for Libraries. Palo Alto, 2004. 299-305.

[4] Shi BL, Zhang L, Wang Y, Chen ZF, Content-Based Chinese antique books retrieval through visual similarity criteria. Journal of Software, 2001,12(9):1336-1342 (in Chinese with English abstract).

[5] Segal B. Grid computing: The European data grid project. In: Proc. of the 2000 IEEE Nuclear Science Symp. and Medical Imaging Conf. Lyon, 2000.

[6] Hoschek W, Jaen-Martinez J, Samar A, Stockinger H, Stockinger K. Data management in an international data grid project. In: Proc. of the 1st IEEE/ACM Int'l Workshop on Grid Computing. Berlin: Springer-Verlag, 2001. 17-20.

[7] Smith J, Gounaris A, Watson P, Paton NW, Fernandes AAA, Sakellariou R. Distributed query processing on the grid. In: Proc. of the 3rd Int'l Workshop on Grid Computing. Berlin: Springer-Verlag, 2002. 279-290.

[8] Yang DH, Li JZ, Zhang WP. Grid-Based join operation. Journal of Computer Research and Development, 2004,41(10):1848-1855 (in Chinese with English abstract).

[9] B?hm C, Berchtold S, Keim DA. Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases. ACM Computing Surveys, 2001,33(3):322-373.

[10] Guttman A. R-Tree: A dynamic index structure for spatial searching. In: Yormark B, ed. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD'84). Boston: ACM Press, 1984. 47-54.

[11] Weber R, Schek HJ, Blott S. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In: Gupta A, Shmueli O, Widom J, eds. Proc. of the 24th Int'l Conf. on Very Large Data Bases (VLDB'98). New York: Morgan Kaufmann Publishers, 1998. 194-205.

[12] Fonseca MJ, Jorge JA. NB-Tree: An indexing structure for content-based retrieval in large databases. In: Proc. of the 8th Int'l Conf. on Database Systems for Advanced Applications. Kyoto: IEEE Computer Society, 2003. 267-274.

[13] Jagadish HV, Ooi BC, Tan KL, Yu C, Zhang R. iDistance: An adaptive B+-tree based indexing method for nearest neighbor search. ACM Trans. on Data Base Systems, 2005,30(2):364-397.

[14] Zhuang YT, Zhang XF, Wu JQ, Lu XQ. Retrieval of Chinese calligraphic character image. In: Aizawa K, Nakamura Y, Satoh S, eds. Proc. of the Pacific Rim Conf. on Multimedia (PCM 2004). Berlin, Heidelberg: Springer-Verlag, 2004. 63-84.

[15] Jagadish HV, Ooi BC, Vu QH, Zhang R, Zhou AY. VBI-Tree: A peer-to-peer framework for supporting multi-dimensional indexing schemes. In: Proc. of the 22nd IEEE Int'l Conf. on Data Engineering (ICDE 2004). New York: IEEE Computer Society Press, 2004.

[16] Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases. In: Jagadish HV, Mumick IS, eds. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD'96). New York: ACM Press, 1996. 103-114.

[17] The cadal project. 2006. http://www.cadal.zju.edu.cn

[18] Cohen S, Guibas L. The earth mover's distance under transformation sets. In: Proc. of the Int'l Conf. on Computer Vision (ICCV'99). New York: IEEE Computer Society Press, 1999. 173-187.

[19] Wu YS, Ding XQ. The Recognition of Chinese Character: Principle, Approach and Implementation. Beijing: Higher Education Press, 1992 (in Chinese).

附中文参考文献:

[4] 施伯乐,张亮,王勇,陈智锋.基于视觉相似性的计算机古籍内容检索方法.软件学报,2001,12(9):1336-1342.

[8] 杨东华,李建中,张文平.基于数据网格环境的连接操作算法.计算机研究与发展,2004,41(10):1848-1855.

[19] 吴佑寿,丁晓青.汉字识别——原理、方法与实现.北京:高等教育出版社,1992.