纯Peer to Peer环境下有效的Top-k查询

何盈捷, 王 珊, 杜小勇

何盈捷, 王 珊, 杜小勇

(北京大学 信息科学技术学院,北京 100871)

作者简介: 何盈捷(1977－),男,浙江诸暨人,博士,主要研究领域为数据库,信息检索;王珊(1944－),女,教授,博士生导师,主要研究领域为高性能数据库,数据仓库,知识工程;杜小勇(1963－),男,教授,博士生导师,主要研究领域为高性能数据库,智能信息检索,知识工程.

联系人: 何盈捷 Phn: +86-10-62519453, E-mail: heyj@ruc.edu.cn, http://www.ruc.edu.cn

Abstract

Most of the existing peer-to-peer (P2P) systems only support simple title-based search, and users cannot search the data based on their content. Top-k query is widely used in the search engine and gains great success. However, Processing top-k query in pure P2P network is very challenging because a P2P system is a dynamic and decentralized system. An efficient hierarchical top-k query processing algorithm based on histogram is proposed. First, a distributed query processing model for top-k query is proposed. It does top-k query in a hierarchical way. Ranking and merging of documents are distributed across the peers, which takes full advantage of the computing resource of the network. Next, a histogram is constructed for each peer according to the top k results returned by the peer, and used to estimate the possible upper bound of the score for the peer. By the histogram information, the most possible peers are selected to send the query, so as to greatly improve the search efficiency. Experimental results show that the top-k query improves the query effectiveness, and the histogram improves the query efficiency.

He YJ, Wang S, Du XY. Efficient Top-k query processing in pure peer-to-peer network. *Journal of Software*, 2005,16(4):540-552.

http://www.jos.org.cn/1000-9825/16/540.htm

摘要

目前大多数的Peer-to-Peer(P2P)系统只支持基于文件标识的搜索,用户不能根据文件的内容进行搜索.Top-k查询被广泛地应用于搜索引擎中,获得了巨大的成功.可是,由于P2P系统是一个动态的、分散的系统,在纯的P2P环境下进行top-k查询是具有挑战性的.提出了一种基于直方图的分层top-k查询算法.首先,采用层次化的方法实现分布式的top-k查询,将结果的合并和排序分散到P2P网络中的各个节点上,充分利用了网络中的资源.其次,根据节点返回的结果为节点构建直方图,利用直方图估计节点可能的分数上限,对节点进行选择,提高了查询效率.实验证明,top-k查询提高了查询效果,而直方图则提高了查询效率.

References:

[1] Napster Home Page. http://www.napster.com/

[2] Gnutella Home Page. http://www.gnutella.com/

[3] Stoica I, Morris R, Karger D, Kaashoek MF, Balakrishnan H. Chord: A scalable peer-to-peer lookup service for internet applications. ACM SIGCOMM Computer Communication Review, 2001,31(4):149-160.

[4] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. Boston: Addison Wesley, 1999. 27-30.

[5] Palmer CR, Steffan JG. Generating network topologies that obey power law. In: Proc. of the GLOBECOM. San Francisco: IEEE, 2000. 434-438. http://citeseer.ist.psu.edu/palmer00generating.html

[6] Ripeanu M. Peer-to-Peer architecture case study: Gnutella network. Technical Report, TR-2001-26, University of Chicago, 2001.

[7] Buckley C. Implementation of the SMART information retrieval system. Technical Report, TR35-686, Cornell University, 1985.

[8] Yu C, Philip G, Meng WY. Distributed top-n query processing with possibly uncooperative local systems. In: Freytag JC, Lockemann PC, Abiteboul S, Carey MJ, Selinger PG, Heuer A, eds. Proc. of the 29th Int'l Conf. on Very Large Data Bases. Berlin: Morgan Kaufmann Publishers, 2003. 117-128.

[9] Gravano L, Garcia-Molina H, Tomasic A. The effectiveness of GLOSS for the text database discovery problem. In: Snodgrass RT, Winslett M, eds. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 1994. 126-137.

[10] Callan JP, Lu ZH, Croft WB. Searching distributed collections with inference networks. In: Proc. of the 18th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Seattle: ACM Press, 1995. 21-28.

[11] Zhang W, Li JZ, Gao H, Pan HW. K-NN query processing in peer-to-peer multimedia database. Computer Science, 2002,29(8):190-193.

[12] Crespo A, Garcia-Molina H. Routing indices for peer-to-peer systems. In: Proc. of the 22nd Int'l Conf. on Distributed Computing Systems. Vienna: IEEE Computer Society, 2002. 23-34. http://dbpubs.stanford.edu:8090/pub/2001-48

[13] Yang B, Garcia-Molina H. Efficient search in peer-to-peer networks. In: Proc. of the 22nd Int'l Conf. on Distributed Computing Systems. Vienna: IEEE Computer Society, 2002. 5-14. http://dbpubs.stanford.edu:8090/pub/2001-47

[14] Cuenca-Acuna FM, Nguyen TD. Text-Based content search and retrieval in ad hoc P2P communities. Technical Report, DCS-TR-483, Department of Computer Science, Rutgers University, 2002.

[15] Tang CQ, Yu ZH, Mahalingam M. pSearch: Information retrieval in structured overlays. ACM SIGCOMM Computer Communication Review, 2003,33(1):89-94.

[16] Ratnasamy S, Francis P, Handley M, Karp RM, Schenker S. A scalable content-addressable network. ACM SIGCOMM Computer Communication Review, 2001,31(4):161-172.

[17] Ling B, Lu ZG, Ng W, Ooi BC, Tan KL, Zhou AY. A content-based resource location mechanism in PeerIS. In: Proc. of the 3rd Int'l Conf. on Web Information Systems Engineering. Singapore: IEEE, 2002. 279-289. http://citeseer.ist.psu.edu/ling02contentbased.html

[18] Ng W, Ooi BC, Tan KL, Wang XY, Ling B, Zhou AY. A novel peer-to-peer system based on self-configuration. Journal of Software, 2003,14(2):237-246 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/237.htm

附中文参考文献：
[11] 张炜,李建中,高宏,潘海为.基于Peer-to-Peer的多媒体数据库K-NN查询处理.计算机科学,2002,29(8):190-193.

[18] 黄维雄,黄铭钧,陈建利,王晓宇,凌波,周傲英.一种基于自配置策略的新型peer-to-peer平台系统.软件学报,2003,14(2):237-246. http://www.jos.org.cn/1000-9825/14/237.htm