大数据存储系统与管理
课程报告

华宇

https://csyhua.github.io/

基本结构

- 数据结构的设计
- 操作流程分析：如何保证和实现所提出的设计目标
- 理论分析(如有false positive和false negative)
- 实验测试的性能：查询延迟，空间开销，错误率等性能指标

提交课程报告

- 提交方式：网上提交到**课程平台**，课程报告标准格式

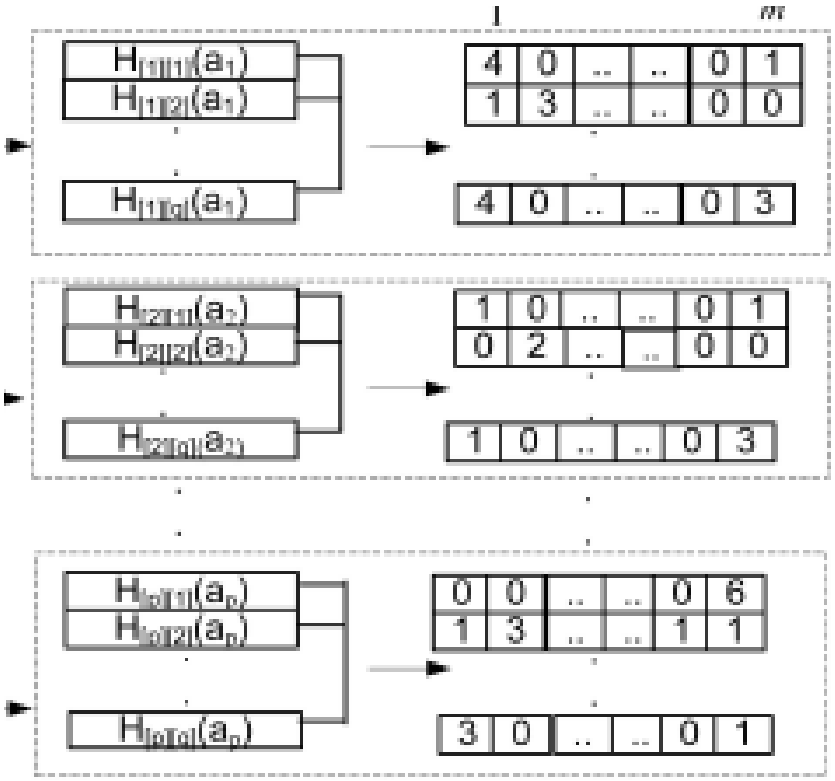
- 电子版本：注明“学号+姓名”，电子版本word文件

- **提交时间：**

2025年6月6日前通过网络提交到**课程平台**，课程报告和实验报告是相同提交方法

选题1：基于Bloom Filter的设计

- 基于Bloom Filter的多维数据属性表示和索引（系数0.8）

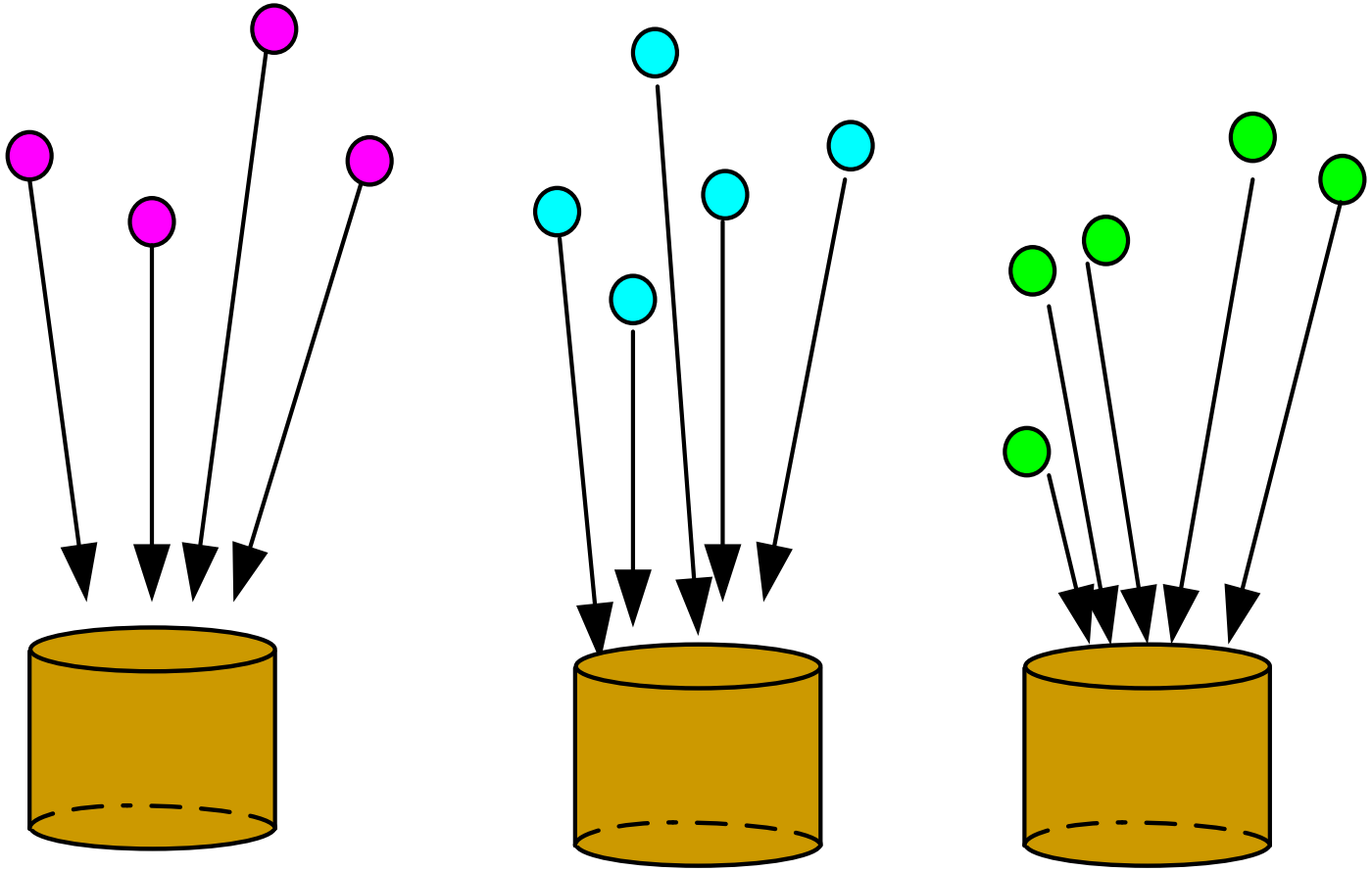


参考文献

- F. Bonomi, M. Mitzenmacher, R. Panigrahy, S. Singh, and G. Varghese, “Beyond Bloom Filters: From Approximate Membership Checks to Approximate State Machines,” Proc. ACM SIGCOMM, 2006.
- Y. Zhu and H. Jiang, “False Rate Analysis of Bloom Filter Replicas in Distributed Systems,” Proc. Int’l Conf. Parallel Processing (ICPP ’06), pp. 255-262, 2006.
- S. Dharmapurikar, P. Krishnamurthy, and D.E. Taylor, “Longest Prefix Matching Using Bloom Filters,” Proc. ACM SIGCOMM, pp. 201-212, 2003.
- L. Fan, P. Cao, J. Almeida, and A. Broder, “Summary Cache: A Scalable Wide-Area Web Cache Sharing Protocol,” IEEE/ACM Trans. Networking, vol. 8, no. 3, pp. 281-293, June 2000.
- B. Xiao and Y. Hua, “Using Parallel Bloom Filters for Multi- Attribute Representation on Network Services,” IEEE Trans. Parallel and Distributed Systems, vol. 21, no. 1, pp. 20-32, Jan. 2010.
- Y. Hua, Y. Zhu, H. Jiang, D. Feng, and L. Tian, “Scalable and Adaptive Metadata Management in Ultra Large-scale File Systems,” Proc. 28th Int’l Conf. Distributed Computing Systems (ICDCS ’08), pp. 403-410, 2008.
- D. Guo, J. Wu, H. Chen, and X. Luo, “Theory and Network Application of Dynamic Bloom Filters,” Proc. IEEE INFOCOM, 2006.

选题2：基于LSH的设计和实现

如何减少LSH的空间开销(系数1)

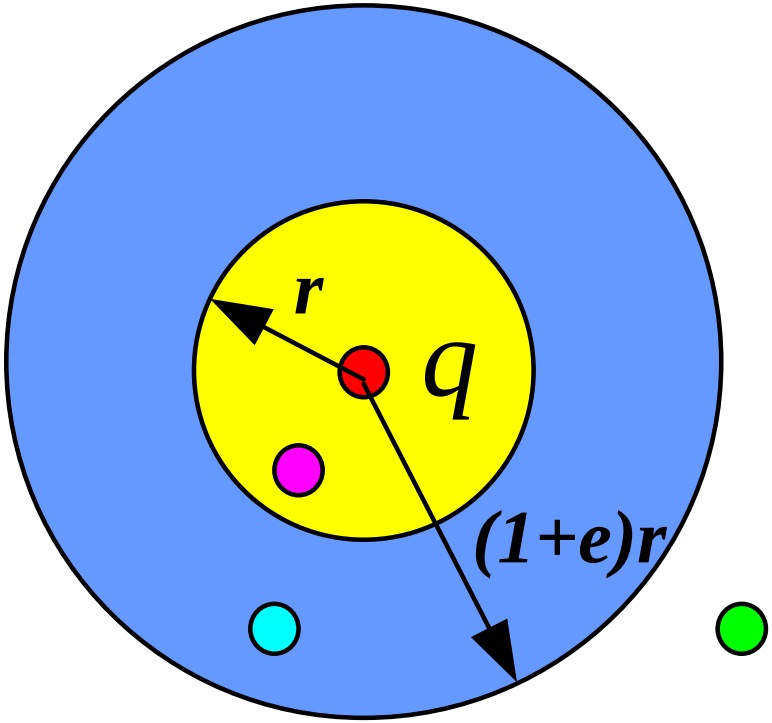


Locality Sensitive Hashing (LSH)

- If $||p,q||_s \leq R$ then $Pr_{\mathbb{H}}[h(p) = h(q)] \geq P_1$,
- If $||p,q||_s > cR$ then $Pr_{\mathbb{H}}[h(p) = h(q)] \leq P_2$.

Near neighbor?

- ***yes***
- ***not sure***
- ***no***

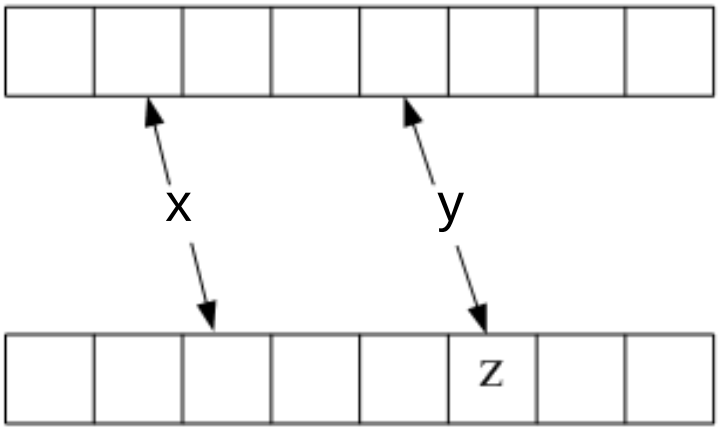


参考文献

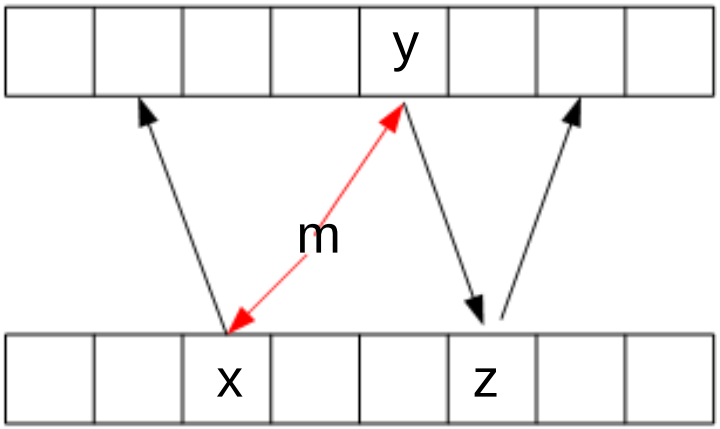
- "Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions" (by Alexandr Andoni and Piotr Indyk). Communications of the ACM, vol. 51, no. 1, 2008, pp. 117-122.
- Q. Lv, W. Josephson, Z. Wang, M. Charikar, and K. Li, “Multi- Probe LSH: Efficient Indexing for High-Dimensional Similarity Search,” Proc. 33rd Int’l Conf. Very Large Data Bases (VLDB ’07), pp. 950-961, 2007.
- Yu Hua, Bin Xiao, Bharadwaj Veeravalli, Dan Feng. "Locality-Sensitive Bloom Filter for Approximate Membership Query", IEEE Transactions on Computers (TC), Vol. 61, No. 6, June 2012, pages: 817-830.
- Yu Hua, Xue Liu, Dan Feng, "Data Similarity-aware Computation Infrastructure for the Cloud", IEEE Transactions on Computers (TC), Vol.63, No.1, January 2014, pages: 3-16.

选题3： Cuckoo-driven Way

如何确定循环，减少cuckoo操作中的无限循环的概率和有效存储(系数1.2)



Insert item x and y



Insert item m

参考文献

- R. Pagh and F. Rodler, “Cuckoo hashing,” Proc. ESA, pp. 121–133, 2001.
- Yu Hua, Hong Jiang, Dan Feng, "FAST: Near Real-time Searchable Data Analytics for the Cloud", Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), November 2014, Pages: 754-765.
- Yu Hua, Bin Xiao, Xue Liu, "NEST: Locality-aware Approximate Query Service for Cloud Computing", Proceedings of the 32nd IEEE International Conference on Computer Communications (INFOCOM), April 2013, pages: 1327-1335.
- Qiuyu Li, Yu Hua, Wenbo He, Dan Feng, Zhenhua Nie, Yuanyuan Sun, "Necklace: An Efficient Cuckoo Hashing Scheme for Cloud Storage Services", Proceedings of IEEE/ACM International Symposium on Quality of Service (IWQoS), 2014.
- B. Fan, D. G. Andersen, and M. Kaminsky, “MemC3: Compact and concurrent memcache with dumber caching and smarter hashing,” Proc. USENIX NSDI, 2013.
- B. Debnath, S. Sengupta, and J. Li, “ChunkStash: speeding up inline storage deduplication using flash memory,” Proc. USENIX ATC, 2010.