Course Form for PKU Summer School International 2018

Course Title	**Compact Data Structures for Big Data** 紧凑数据结构与大数据
Teacher	CHEN Shigang
First day of classes	July 15, 2018
Last day of classes	July 22, 2018
Course Credit	2 credits

Course Description

Objective

This course covers compact data structures, algorithms, probabilistic methods, and statistical tools for handling big data, particularly in the setting of high-speed networks. There is hardly any other data set whose size can rival the big network data that flows on the Internet. Analyzing this big data is extremely useful in improving network performance, user experience and cybersecurity. However, storing such data for analysis is impossible. With this background, the course offers a series of compact data structures and their theoretical analysis that are developed over the last three decades, with increasing capabilities of making big data small for storage and quick access of data properties. The offered data structures and their associated algorithms are broadly classified into two categories: (1) counting and cardinality algorithms, including probabilistic counting, bitmap algorithms, FM sketch, hyperloglog sketch, virtual bitmap, virutal FM sketch, virtual hyperloglog, countMin, counter braids, randomized counter sharing, and virtual counters, and (2) membership lookup and classification, including Bloom filters, counting Bloom filters, Bloomier filters, blocked Bloom filters, multi-set filters, and multi-hashing tables. The students will be exposed to not only these data structures and algorithms, but also their applications in Internet traffic measurement, cybersecurity, as well as applications beyond the network context. The students are expected to learn the compact data structures and algorithms through lectures, and implement a selected subset with experiments over real network data.

Pre-requisites /Target audience

Data Structures and Algorithm, Computer Networks
Senior undergraduate students and graduate students

Proceeding of the Course

Data structure

Assignments (essay or other forms)

Reading, Assignment and Programming

Evaluation Details

Attendance and Reading: 30%
Programming Project: 40%
Exam: 30%

Text Books and Reading Materials

1. Shigang Chen, Min Chen, Qingjun Xiao. Traffic Measurement for Big Network Data. Springer, ISBN 978-3-319-47340-6, 2016.
2. Tao Li, Shigang Chen. Traffic Measurement on the Internet. Springer, ISBN: 978-1-4614-4850-1, 2012.
3. Kyu-young Whang, Brad T. Vander-Zanden, Howard M. Taylor. A Linear-Time Probabilistic Counting Algorithm for Database Applications. ACM Transactions on Database Systems, Vol. 15, No. 2, 1990.
4. Cristian Estan, George Varghese, Mike Fisk. Bitmap Algorithms for Counting Active Flows on High Speed Links. ACM Internet Measurement Conference, 2003.
5. Peter Lieven, Bj¨orn Scheuermann. High-Speed Per-Flow Traffic Measurement with Probabilistic Multiplicity Counting. IEEE INFOCOM, 2010.
6. Philippe Flajolet, Éric Fusy, Olivier Gandouet and Frédéric Meunier. HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm. Conference on Analysis of Algorithms, 2007.
7. Qingjun Xiao, Shigang Chen, Min Chen, Yibei Ling. Hyper-Compact Virtual Estimators for Big Network Data Based on Register Sharing. ACM SIGMETRICS, 2015.
8. Yi Lu, Andrea Montanari, Balaji Probhakar, Sarang Dharmapurikar, and Abdul Kabbani. Counter Braids: A Novel Counter Architecture for Per-Flow Measurement. ACM SIGMETRICS, 2008.
9. Andrei Brodery, Michael Mitzenmacherz. Network Applications of Bloom Filters: A Survey, Internet mathematics, 2004.
10. Fang Hao, Murali Kodialam, T. V. Lakshman, and Haoyu Song. Fast Dynamic Multiple-Set Membership Testing Using Combinatorial Bloom Filters. IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 20, NO. 1, 2012.

Academic Integrity (If necessary)

CLASS SCHEDULE
(Subject to adjustment)

Session 1： *Probabilistic Counting*	Date:7/28

【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) basic methods and analytical tools for designing sketches
【Questions】
【Readings, Websites or Video Clips】 Kyu-young Whang, Brad T. Vander-Zanden, Howard M. Taylor. A Linear-Time Probabilistic Counting Algorithm for Database Applications. ACM Transactions on Database Systems, Vol. 15, No. 2, 1990.
【Assignments for this session (if any)】

Session 2：*Bitmap Algorithms*	Date:7/29

【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) various bitmap algorithms for per-flow cardinality estimation
【Questions】
【Readings, Websites or Video Clips】 Cristian Estan, George Varghese, Mike Fisk. Bitmap Algorithms for Counting Active Flows on High Speed Links. ACM Internet Measurement Conference, 2003.
【Assignments for this session (if any)】

Session 3：*FM Sketch*	Date:7/29

【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) operations and properties of FM sketches for per-flow cardinality estimation
【Questions】
【Readings, Websites or Video Clips】 Peter Lieven, Bj¨orn Scheuermann. High-Speed Per-Flow Traffic Measurement with Probabilistic Multiplicity Counting. IEEE INFOCOM, 2010.

【Assignments for this session (if any)】

Session 4：*Hyperloglog Sketch*	Date:7/30
【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) the state-of-the-art hyperloglog sketches for big data	
【Questions】	
【Readings, Websites or Video Clips】 Philippe Flajolet, Éric Fusy, Olivier Gandouet and Frédéric Meunier. HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm. Conference on Analysis of Algorithms, 2007.	
【Assignments for this session (if any)】	
Session 5：*Virtual Bitmap*	Date:7/30
【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) the basic methods and analytical tools for virtualized data structures	
【Questions】	
【Readings, Websites or Video Clips】	
【Assignments for this session (if any)】	
Session 6：*Virtual FM Sketch*	Date:7/31
【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) different ways of virtualizing FM sketches	

【Questions】
【Readings, Websites or Video Clips】 Peter Lieven, Bj¨orn Scheuermann. High-Speed Per-Flow Traffic Measurement with Probabilistic Multiplicity Counting. IEEE INFOCOM, 2010.
【Assignments for this session (if any)】

Session 7：*Virtual Hyperloglog*	Date:7/31
【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) virtualizing hyperloglog	
【Questions】	
【Readings, Websites or Video Clips】 Qingjun Xiao, Shigang Chen, Min Chen, Yibei Ling. Hyper-Compact Virtual Estimators for Big Network Data Based on Register Sharing. ACM SIGMETRICS, 2015.	
【Assignments for this session (if any)】	
Session 8：*CounterMin*	Date:8/1
【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) simple yet powerful countMin tools for summarizing big data	
【Questions】	
【Readings, Websites or Video Clips】	
【Assignments for this session (if any)】	

Session 9：*Counter Braids*	Date:8/1
【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) extending countMin in two dimensions for better performance	
【Questions】	
【Readings, Websites or Video Clips】 Yi Lu, Andrea Montanari, Balaji Probhakar, Sarang Dharmapurikar, and Abdul Kabbani. Counter Braids: A Novel Counter Architecture for Per-Flow Measurement. ACM SIGMETRICS, 2008.	
【Assignments for this session (if any)】	
Session 10：*Randomized Counter Sharing and Virtual Counters*	Date:8/2
【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) virtualizing counters	
【Questions】	
【Readings, Websites or Video Clips】 Peter Lieven, Bj¨orn Scheuermann. High-Speed Per-Flow Traffic Measurement with Probabilistic Multiplicity Counting. IEEE INFOCOM, 2010.	
【Assignments for this session (if any)】	
Session 11： *Bloom Filters*	Date:8/2
【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) the basic ideas and analytical methods for Bloom filters	
【Questions】	

【Readings, Websites or Video Clips】 Andrei Brodery, Michael Mitzenmacherz. Network Applications of Bloom Filters: A Survey, Internet mathematics, 2004.
【Assignments for this session (if any)】

Session 12： *Counting Bloom Filters*	Date:8/3

【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) augmenting Bloom filters with the ability of deletion
【Questions】
【Readings, Websites or Video Clips】
【Assignments for this session (if any)】

Session 13：*Bloomier Filters*	Date:8/3

【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) Augmenting Bloom filters with the ability of value lookup
【Questions】
【Readings, Websites or Video Clips】
【Assignments for this session (if any)】

Session 14：*Blocked Bloom Filters*	Date:8/3

【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) A different design of Bloom filters
【Questions】
【Readings, Websites or Video Clips】 Andrei Brodery, Michael Mitzenmacherz. Network Applications of Bloom Filters: A Survey, Internet mathematics, 2004.
【Assignments for this session (if any)】

Session 15： *Multiset Bloom Filters*	Date:8/4
【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) variants of Bloom filters for multiset lookup	
【Questions】	
【Readings, Websites or Video Clips】 Fang Hao, Murali Kodialam, T. V. Lakshman, and Haoyu Song. Fast Dynamic Multiple-Set Membership Testing Using Combinatorial Bloom Filters. IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 20, NO. 1, 2012.	
【Assignments for this session (if any)】	
Session 16： *Multi-hash Tables*	Date:8/4
【Description of the Session】(purpose, requirements, class and presentations scheduling, etc.) using multiple hash functions to improve the space efficiency of hash tables	
【Questions】	

【Readings, Websites or Video Clips】
【Assignments for this session (if any)】