CS466 Project Proposal

Mingkun Ni, Yuanhao Zhang
{m8ni, y2384zha}@uwaterloo.ca
University of Waterloo
Waterloo, ON, Canada

July 9, 2020

Background

Hash has been a research question with a long history. The concept of hash is first used in a memo of Hans Peter Luhn in 1953 (Donald K.,2000). After the analysis of the linear probing algorithm by Donald Knuth, the analysis of hashing function becomes famous, and many talented researchers have put efforts into the development of related fields. Nowadays, hash has become not only a well-development research field but also a popular tool that is widely used in programming. Because of its wide utility, we consider it as a focus of our project.

After reading through some related research papers, we decide to look into one of them, titled *Ball and bins: smaller hash families and faster evaluation*. This paper introduces two new constructions that we could guarantee $O(\log n/\log\log n)$ maximum load when throw n balls into n with either a smaller description length or a faster calculation time. It is well-known that, with high probability, a $O(\log n/\log\log n)$-wise independent hash family would guarantee max load of $O(\log n/\log\log n)$. Such $O(\log n/\log\log n)$-wise independent function can be described by $O(\log^2 n/\log\log n)$ bits, which already yields a dramatic improvement over a truly random function. This paper aims to find an even smaller description length or a faster calculation time of the function while maintaining the guarantee of $O(\log n/\log\log n)$ max load. The special part of this research, specifically about the first construction, is that it constructs an innovative structure of a multi-layer random graph. With such construction of multi-layer graph, each layer is considered as a different hash process with different input and output sizes, which is the number of bins in each layer. Thus, we want to look deep in the first construction of this research and elaborate more details.

Plan

This research paper will contain the following 4 main sections:

Section One is the introduction of the research problem including the reasons that we chose this topic, background information on the topic as well as prior work contributed by other researchers. In the introduction, we will be talking about how truly random hashing functions are impractical in various applications due to the space complexity. We would then be explaining how pair-wise independent hashing functions are introduced with a much more feasible description length and space complexity. Then, we will give a brief introduction about the new construction, which guarantees the same max load while reducing the description length of the function.

In Section Two, there will be two subsections. In Subsection One, we will first prove the well-known fact: with high probability, a $O(\log n/\log\log n)$-wise independent hash family would guarantee max load of $O(\log n/\log\log n)$. Also, we will show that the description length of such construction is $O(\log^2 n/\log\log n)$. In Subsection Two, we will list the tools that our formal proof in section 3 will need including some lemmas and theorems as well as their proofs. For example, we will be proving the property of a *2k-wise δ-dependent* random variable, which is an essential step of our formal proof in Section Three.

Section Three will be the most essential part of this research: it will contain the formal introduction of our construction and formal proof of why this construction works. This section will be split into multiple subsections as well. It will first give a formal description of the construction, and we will be analyzing the construction in which we will be essentially explaining the interpretation of such construction. Finally, we will prove that the construction guarantees the description length of hashing functions.

In the end, Section 4 will introduce some extensional uses of this construction. It will also analyze this construction with correspondence to the trade-off it makes to obtain the smaller description length of the function.

We plan to finish reading and collecting information about the construction and related tools before 24th July, and we will start writing our research paper right after that. We expect to finish writing the first draft before 3rd Aug and the final draft before 14th Aug.

References

[1] L. E. Celis, O. Reingold, G. Segev and U. Wieder, ”Balls and Bins: Smaller Hash Families and Faster Evaluation,” 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science, Palm Springs, CA, 2011, pp. 599-608, doi: 10.1109/FOCS.2011.49.

[2] A. Pagh and R. Pagh. Uniform hashing in constant time and optimal space. SIAM Journal on Computing, 38(1):8596, 2008.

[3] J. Naor and M. Naor. Small-bias probability spaces: Efficient constructions and applications. SIAM Journal on Computing, 22(4):838856, 1993.

[4] M. Dietzfelbinger and F. Meyer auf der Heide. A new universal class of hash functions and dynamic hashing in real time. In Proceedings of the 17th International Colloquium on Automata, Languages and Programming, pages 619, 1990.

[5] N. Alon, M. Dietzfelbinger, P. B. Miltersen, E. Petrank, and G. Tardos. Linear hash functions. Journal of the ACM, 46(5):667683, 1999.

[6] N. Alon, O. Goldreich, J. Hastad, and R. Peralta. Simple construction of almost kwise independent random variables. Random Structures and Algorithms, 3(3):289304, 1992.

[7] O. Goldreich and A. Wigderson. Tiny families of functions with random properties: A quality-size trade-off for hashing. Random Structures and Algorithms, 11(4):315343, 1997.