FORMAL VERIFICATION OF THE HEAVY HITTER PROBLEM

Ghassen Helali, Sofiene Tahar

Department of Electrical and Computer Engineering, Concordia University
Montreal, Quebec, H3G 1M8, Canada
Email: {helali,tahar}@encs.concordia.ca

Osman Hasan

SEECS, National University of Sciences and Technology
Sector H-12, Islamabad, Pakistan
Email:osman.hasan@seecs.nust.edu.pk

ABSTRACT

The heavy hitter problem is used to assess the frequency of occurrence of an element in a given data stream. It is one of the most widely used combinatorial tools in many safety-critical domains including medicine, telecommunications and stock exchange markets. Traditionally, the heavy hitter problem is analyzed using paper-and-pencil proofs, simulation or computer algebra systems. These techniques are informal and thus may result in an inaccurate analysis, which poses a serious threat to the reliability of the underlying applications of the problem. To overcome this limitation, we present a formal probabilistic analysis approach for the heavy hitter problem using a higher-order-logic theorem prover (HOL). The paper presents the higher-order-logic model of an algorithm for the heavy hitter problem. This model is then utilized to formally verify some interesting probabilistic and statistical properties associated with the heavy hitter problem in HOL.

Index Terms– Higher-order Logic, Probability Theory, Theorem Proving, Formal Verification, Heavy Hitter Problem.

1. INTRODUCTION

Given a $Universe\ U$, a data stream of length n and a parameter $\lambda \in [0,1]$, the *Heavy Hitter problem* [3] is to find the λ-Heavy Hitter list which contains the elements of the Universe that occur in the data stream at least λn times. The algorithms for the Heavy Hitter problem are widely used to identify frequently encountered items of streams in a compact way. For example, they have been used to identify heavily traded stocks in streams of financial transactions, to detect viruses spread in networks [17], to monitor network traffic for statistical data collection [4] and to detect Distributed Denial of Services (DDoS) [16].

An algorithm for the heavy hitter problem is as follows:

Input: the frequency λ, a data stream DS and a Universe U with length n
Output: the list L of elements from U occurring at least $\lambda * n$ times in DS

$L \leftarrow [\,]$
for $i = 1 \rightarrow n$ **do**
  if $freq(DS[i]) \geq \lambda$ **then**
    $L \leftarrow L$ ***INSERT*** $DS[i]$
  end if
end for

The major goal of analyzing such a problem is to predict the behavior of some precise elements in a data stream. For example, controlling specific transactions in a financial flow. The biggest challenge in this analysis is the unpredictable nature of the input data stream. Therefore, probabilistic techniques are used for the analysis where the algorithm is modeled by an appropriate random variable.

Due to its extensive usage, the heavy hitter problem and some of its variants have been extensively analyzed using probabilistic techniques based on paper-and-pencil proof methods (e.g., [14]), computer simulations (e.g., [15]), and computer algebra systems (e.g., [12]). The traditional paper-and-pencil based proof techniques always involve some risk of an erroneous analysis due to the human-error factor. Computer simulation cannot guarantee correctness since the fundamental idea in this approach is to approximately answer a query by analyzing a large number of samples. Moreover, the random variables are usually modeled using pseudo random number generators in simulation based analysis, which further introduces some approximations. Finally, computer algebra systems also fail to provide precise results because they are constructed using extremely complicated algorithms, which are likely to contain bugs.

The accuracy of the analysis for the heavy hitter problem has become imperative these days because of its extensive usage in highly sensitive applications in areas like medicine and security. Thus, more reliable analysis techniques than the ones discussed above are required. Higher-order-logic theorem proving [5] is capable of conducting precise analysis and therefore can fulfill the above mentioned requirement. Higher-order logic is a system of deduction with a precise semantics and is expressive enough to be used for the precise specification of almost all classical mathematics theories.

2012 25th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)
978-1-4673-1433-6/12/$31.00 ©2012 IEEE

The foremost requirements for conducting the probabilistic analysis of the heavy hitter problem in a higher-order-logic theorem prover are (i) to be able to formalize random variables in higher-order logic, which in turn can be used to formally model the random input behavior of the algorithm for the heavy hitter problem, and (ii) to be able to formally express and verify probabilistic and statistical properties in order to check the interesting performance characteristics in a theorem prover. There have been some significant developments related to the above mentioned criteria in the last few years. Namely, the extended real number based formalization of measure, probability and Lebesgue integration theory [11]. In this paper, we basically build upon these foundations. We first formalize the above algorithm in higher-order-logic along with the random input behavior. This is followed by the formal verification of an interesting performance characteristic, i.e., Chebychev's inequality based bound on the probability of identifying a heavy hitter, within the sound environment of a theorem prover. The analysis results can be claimed to be 100% precise, which is an achievement that has not been reported in the open literature so far.

2. RELATED WORK

The foremost requirement for conducting the formal probabilistic analysis of the heavy hitter problem in a theorem prover is to have access to a higher-order-logic formalization of probability fundamentals. Several formalizations of the probability theory have been reported in the open literature. Coble [1] formalized the main concepts of Lebesgue integration and further used these fundamentals to formalize some information theory in HOL. However, Coble's formalization of Lebesgue integral can only consider finite-valued measures, functions and integrals. Mhamdi [11] generalized Hurd and Coble's work by introducing a Borel space in HOL. He defined the extended real numbers (real numbers including $\pm\infty$) and used them to formalize measure, Lebesgue, probability and information Theories. This way, the monotonicity of the Lebesgue integral could be proved even for the non-integrable functions, and also the convergence theorem for non convergent sequences. On similar lines, Hölzl [10] also formalized the borel σ-algebra as well as its topology concepts in Isabelle/HOL. He used that topology space to formalize measure, probability and Lebesgue integration theory and some of their useful properties. We utilize Mhamdi's work in this paper for formally analyzing the heavy hitter problem. The prime motivation for this choice is the completeness of the work and its availability in the HOL theorem prover, with which we have prior experience.

In the area of higher-order-logic theorem proving based formal probabilistic analysis of algorithms, Hasan analyzed the algorithm for the Coupon Collectors problem [7] and also developed a methodology to analyze the expected time complexity of algorithms [8]. The proposed analysis of the heavy hitter problem is based upon the work of Mhamdi [11] and is thus more general. Besides being the first formal analysis of the famous heavy hitter problem, to the best of our knowledge, the presented work is also the first practical application of Mhamdi's formalization of probability theory and thus demonstrates its usefulness for analyzing real-world problems.

3. FORMALIZATION OF THE HEAVY HITTER PROBLEM

As mentioned in [3], *The ϵ-relaxed λ-heavy hitter problem is to find a set $H \subseteq U$ such that all the elements of H appears at least λn times*

The Heavy Hitter Problem can be formalized in HOL by modeling the sample set of elements and the data stream as lists. Then we model a function, `freq`, that returns the frequency of an element in a list which is defined below,

Definition 1: *Frequency of an Element in a List*

```
⊢ ∀ e L. freq e L =
  ((LENGTH(FILTER(λr.r = e) L)))/
  (((LENGTH L)))
```

where `LENGTH` returns the length of a list, and `FILTER` returns a filtered list out of its argument list with elements that satisfy the given condition. The above function will be required later, to report the list of the α-heavy hitter elements. Next, we model another function, `HeavyHitter_lst`, which takes as parameters two lists and a real value and returns the list of heavy hitters corresponding to the algorithm of heavy hitter problem.

Definition 2: *The α-Heavy Hitter Algorithm*

```
⊢ ∀ L M α.  HeavyHitter_lst L M α =
  FILTER(λr.α <= (freq (EL r L)M))L
```

So far, our development has been based on deterministic functions. We now introduce the randomness in our models using appropriate random variables and our probabilistic analysis of the algorithm for the heavy hitter problem would be based upon the characteristics of these random variables. In our analysis, we need a Bernoulli variable X, with outcomes 1 or 0. In order to formalize this random behavior, we have to define a probability space that has the set $\{0, 1\}$ as its space and power set, `POW` $\{0, 1\}$, as the events space and the probability measure will be a new function that returns pr if the set in parameter implements the fact that $f(x)$, which refers to our random variable in this case, is equal to 1, and returns $1 - pr$ otherwise. The corresponding probability measure is defined in HOL as follows:

Definition 3: *The Probability Measure of Bernoulli Random Variable*

```
⊢ ∀ g pr.  mu g pr =
  (λa.if (a = x|g x = 1) then pr
   else (1 - pr))
```

The new probability space is formalized as follows:

Definition 4: *The Heavy Hitter Probability Space*

```
⊢ ∀ g pr.  HH_prob_space pr g =
   ({0;1},POW {0;1},mu g pr)
```

Finally, the new random variable will be modeled as

Definition 5: *The Heavy Hitter Random Variable*

```
⊢ ∀ X pb.  HH_rv X pb = random_variable X
   (HH_prob_space pb X) Borel
```

This completes the formal specification of the algorithm for the heavy hitter problem in higher-order logic. We will use these definitions to formally reason about some interesting probabilistic and statistical properties of this algorithm in the next section.

4. FORMAL ANALYSIS OF THE HEAVY HITTER ALGORITHM

We have utilized the HOL theorem prover to formally verify the desired characteristics of the heavy hitter problem. We begin by first verifying the probability mass function (PMF) of the heavy hitter random variable (or the Bernoulli random variable). Since the element of the sample is taken uniformly at random, $Pr[X_i = 1] = \frac{\lambda n}{n} = \lambda$, which is represented in the HOL specification by pr. This theorem can be easily proved in HOL using only the definition of the new random variable as follows:

Theorem 1: *PMF of Heavy Hitter Random Variable (RV)*

```
⊢ ∀i X. HH_rv (X i) pr ⇒
  prob (HH_prob_space pr (X i))
   {x | X i x = 1} = pr
```

Next, we verify the expectation of the heavy hitter random variable as the following theorem in HOL.

Theorem 2: *Expectation of the Heavy Hitter RV*

```
⊢ ∀s.  (HH_rv (X i') pr)∧
  (∀i.  i ∈ s ⇒
   (X i ∈ measurable
    (m_space (HH_prob_space pr (X i')),
     measurable_sets
    (HH_prob_space pr (X i'))) Borel))) ⇒
  expectation (HH_prob_space pr (X i'))
   (λx.  ∑ (λi.  X i x) s) =
    pr * (CARD s)
```

The assumptions used above refer respectively to our new random variable defined in Definition 5 and each random variable is measurable with respect to our new probability space cited in Definition 4. The HOL function `CARD` is the mathematical cardinality operator. The formal reasoning for the above theorem was based on the definition of the expectation and some real analysis properties, i.e, $\sum_{(i=1)}^{n} a = na$.

Let's choose the cardinality of s as $cs = \frac{4}{\delta.\epsilon^2}$. We would like to prove, using our new formalizations as well as the Chebyshev's inequality and some real analysis, the property saying that: *the probability of reporting x is at least* $(1 - \delta)$.

$$\mathbf{Pr}[\textstyle\sum_i X_i > cs * (\lambda - \epsilon/2)] \geq 1 - \delta$$

This can be formalized in HOL as follows

Theorem 3: *The Upper Bound of the Probability of Identifying a Heavy Hitter*

```
⊢(∀ e s pr.  (FINITE s) ∧
  HH_rv (X i') pr ∧
  (∀i.  i ∈ s ⇒
   X i ∈ measurable
     (m_space (HH_prob_space pr (X i')),
     measurable_sets
     (HH_prob_space pr (X i'))) (Borel)) ⇒
  (1 - 4/(cs×(e²))) <=
  prob (HH_prob_space pr (X i'))
   {x|x ∈ p_space
     (HH_prob_space pr (X i')) ∧
    cs×(α - (e/2)<(∑ (λi.  X i x) s))}
```

In order to verify this result, we procede by dividing our main goal into a number of subgoals. The first subgoal is to verify the following relationship

$\{x \mid \mathbf{E}[\sum_i X_i] - \sum_i X_i \ < \ (\frac{e}{2}) \times cs\}$ =
$\{x \mid \mathbf{E}[\sum_i X_i] \ - \ (\frac{e}{2}) \times cs) \ < \ \sum_i X_i\}$.

The formal reasoning for the above subgoal was based on Theorem 2 and some properties of the inequalities. The following subgoal extends the previous result by applying the probability measure and then using the available property of the probability of the complementary events.

$Pr[\mathbf{E}[\sum_i X_i] - \sum_i X_i \ < \ (\frac{e}{2}) \times cs]$ =
$1 \ - \ Pr[\mathbf{E}[\sum_i X_i] \ - \ \sum_i X_i \ \geq \ (\frac{e}{2}) \times cs)]$.

The next subgoal is to verify the formally probabilistic relationship

Pr[(E[$\sum X_i$] - $\sum X_i$)$\geq$ ($\frac{e}{2}$) $\times$ cs] $\leq$
Pr[|$\sum X_i$) - (E[$\sum X_i$]|$\geq$ ($\frac{e}{2}$) $\times$ cs]

which is expressed in HOL as follows

```
⊢ ∀ e s pr.  (FINITE s) ∧
  (HH_rv (X i') pr) ⇒
  (prob (HH_prob_space pr (X i'))
   {x|x IN p_space
     (HH_prob_space pr (X i')) ∧
    (e/2) × cs ≤
    (expectation (HH_prob_space pr
     (X i')) (λx.  ∑_{i∈s} (X i x))) -
    (∑_{i∈s} (X i x))} ≤
```

```
prob (HH_prob_space pr (X i'))
 {x|x ∈ p_space
  (HH_prob_space pr (X i')) ∧
 (e/2) × cs ≤
 abs((∑_{i∈s} (X i x)) −
  (expectation (HH_prob_space pr
 (X i')) (λx.  (∑_{i∈s} X i x))))})
```

The proof of the last subgoal is based on the property, *if A1⊆ A2 then Pr[A1] ≤ Pr[A2]* and real analysis. Now Theorem 3 can be verified based on the above subgoals and the formally verified *Chebychev*'s inequality [11] in addition to some real analysis.

The proof script of the formalization of the Heavy Hitter problem presented in this section consists of approximately 530 lines of HOL code. A detailed explanation of the proof can be found in [9]. It is worthwhile to mention here that the results presented in this section are not something that is new and they have been known for quite some time now. The real contribution of the paper lies in demonstrating the ability to achieve these results precisely using a computer based tool.

5. CONCLUSIONS

In this paper, we utilized the mathematical probability theory formalized in the higher-order-logic theorem prover HOL to conduct the probabilistic analysis of the Heavy Hitter problem. The main idea behind our approach is to construct a higher-order-logic model of the algorithm for the Heavy Hitter problem, along with its random components and to verify interesting probabilistic characteristics in a theorem prover. We utilize this approach for analyzing the probability bounds for the Heavy Hitter problem using the Chebychev's inequality. Because of the formal nature of the model and the soundness of the mechanical theorem prover, the analysis is guaranteed to be free of approximation and precision errors, which is a novelty that, to the best of our knowledge, has not been achieved by other probabilistic analysis approaches.

The proposed higher-order-logic theorem proving based probabilistic analysis approach can be applied for the algorithm analysis of many other problems, such as the *hat-check* problem [6], the *hiring* problem [2], the *balls and bins* problem [2], the *longest streak of heads* problem [2], the *on-line hiring* problem [2], the *Chinese appetizer* problem [6] and the *Quicksort* algorithm [13].

6. REFERENCES

[1] A. R. Coble. *Anonymity, Information, and Machine-Assisted Proof*. PhD thesis, University of Cambridge, Cambridge, UK, 2010.

[2] T.H. Cormen, C.E. Leiserson, R.L.Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, 2001.

[3] C.Sohler. The Heavy Hitter problem. In *Streaming Algorithms*, pages 25–28.

[4] S.Guo F. Wang, B. Gong and X.Wang. Monitoring Heavy-Hitter Flows in High-Speed Network Concurrently. In *Network and System Security*, pages 160–166. IEEE Computer Society, 2010.

[5] M.J.C. Gordon. *Mechanizing Programming Logics in Higher Order Logic*. Technical Report. University of Cambridge, Computer Laboratory, 1988.

[6] C.M. Grinstead and J.L. Snell. *Introduction to Probability*. American Mathematical Society, 1997.

[7] O. Hasan. *Formal Probabilistic Analysis Using Theorem Proving*. PhD thesis, Concordia University, Montreal, Q.C., Canada, 2008.

[8] O. Hasan and S. Tahar. Formally Analyzing Expected Time Complexity of Algorithms Using Theorem Proving. *Journal of Computer Science Technolgy*, 25(6):1305–1320, 2010.

[9] G. Helali. *Formalization of the Heavy Hitter Problem in HOL*. Technical Report, Concordia University, Montreal, Quebec, Canada. January 2012, http://hvg.ece.concordia.ca/Publications/TECH_REP/HHP_TR12/.

[10] J. Hölzl and A. Heller. Three Chapters of Measure Mheory in Isabelle/HOL. In *Interactive Theorem Proving*, volume 6898 of *LNCS*, pages 135–151, 2011.

[11] T. Mhamdi, O. Hasan, and S. Tahar. Formalization of entropy measures in HOL. In *Interactive Theorem Proving*, volume 6898 of *LNCS*, pages 233–248, 2011.

[12] F. P. Miller, A. F. Vandome, and M. B. John. *Computer Algebra System*. VDM Verlag Dr. Mueller e.K., 2010.

[13] M. Mitzenmacher and E. Upfal. *Probability and Computing*. Cambridge University Press, 2005.

[14] M. Mitzenmacherl and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

[15] S. M. Ross. *Simulation*. Statistical Modeling and Decision Science. Elsevier Academic Press, 2006.

[16] V. Sekar and J. Van Der Merwe. Lads: Large-Scale Automated DDoS Detection System. In *Proc. of USENIX ATC*, pages 171–184, 2006.

[17] M. Thottan, G. Liu, and C. Ji. Anomaly detection approaches for communication networks. *Algorithms for Next Generation Networks*, pages 239–261, 2010.