

Formally verified algorithms and data structures in Coq: concepts and techniques

(Formalnie zweryfikowane algorytmy i struktury danych w Coqu: koncepcje i
techniki)

Wojciech Kołowski

Praca magisterska

Promotor: oficjalnie nikt

Uniwersytet Wrocławski
Wydział Matematyki i Informatyki
Instytut Informatyki

Czerwiec '20 chyba że koronawirus

Abstract

We discuss how to design, implement, specify and verify functional algorithms and data structures, concentrating on formal proofs rather than asymptotic complexity or actual performance. We present concepts and techniques, both of which often rely on one key principle – the reification and representation, using Coq’s powerful type system, of something which in the classical-imperative approach is intangible, like the flow of information in a proof or the shape of a function’s recursion. We illustrate our approach using rich examples and case studies.

Omawiamy sposoby projektowania, implementowania, specyfikowania i weryfikowania funkcyjnych algorytmów i struktur danych, skupiając się bardziej na dowodach formalnych niż na asymptotycznej złożoności czy faktycznym czasie działania. Prezentujemy koncepty i techniki, obie często opierające na jednej kluczowej zasadzie – reifikacji i reprezentacji, za pomocą potężnego systemu typów Coqa, czegoś co w klasycznym, imperatywnym podejściu jest nieuchwytne, jak przepływ informacji w dowodzie czy kształt rekursji funkcji. Nasze podejście bogato ilustrujemy przykładami i studiami przypadku.

Contents

1	Introduction	7
1.1	The overarching paradigm	7
1.2	Two flavours of algorithms	9
1.3	The many-worlds interpretation of the algorithmic business	11
1.4	An ultra short literature review	13
1.5	Way of the Coq	13
2	A quick <i>tour de</i> sort	15
3	A man, a plan, a canal – MSc thesis	17
3.1	Design	17
3.2	Techniques	17
3.3	Topics	17
	Bibliography	19

Chapter 1

Introduction

The title of this thesis is “Formally verified functional algorithms and data structures in Coq: concepts and techniques”. In the Introduction, we take time to carefully explain what we mean by each of these phrases:

- In section 1.1 we look at the social context that gives rise to an interesting misconception about “algorithms and data structures”.
- In section 1.2 we explain “functional” programming by contrasting it with imperative programming.
- In section 1.3 we lay out the unity of the “formally verified” world by comparing it with the many worlds which are the arenas of the usual algorithmic struggles.
- In section 1.4 we do an ultra short literature review and discuss how our “concepts and techniques” differ from these that can be found there.
- In section 1.5 we show how working “in Coq” looks like and describe some theory and principles Coq is based on.

In the remaining chapters we describe the actual concepts and techniques, using well-known and well-studied examples like quicksort, merge sort or binary heaps:

- TODO

1.1 The overarching paradigm

The Free Dictionary says [1] that an algorithm is

A finite set of unambiguous instructions that, given some set of initial conditions, can be performed in a prescribed sequence to achieve a certain goal and that has a recognizable set of end conditions.

The purpose of this entry is to explain the concept to a lay person, but it likely sounds just about right to the imperative programmer’s ear too. To a functional ear, however, talking about sequences of instructions most certainly sounds as unfunctional as it possibly could. It is no surprise then that some people wonder if it is even possible for algorithms to “exist” in a functional programming language, as exemplified by this StackOverflow question [2]. The poor soul asking this question had strongly associated algorithms with imperative languages in his head, even though functional languages have their roots in lambda calculus, a formal system invented precisely to capture what an algorithm is.

This situation is not uncommon and rather easy to explain. *Imperative* algorithms and data structures¹ form one of the oldest, biggest, most widespread and prestigious fields of theoretical computer science. They are taught to every student in every computer science programme at every university. There’s a huge amount of books and textbooks, with classics such as [3] [4] known to pretty much everybody, at least by title. There’s an even huger and still growing mass of research articles published in journals and conferences and I’m pretty sure there are at least some (imperative) algorithm researchers at every computer science department in existence.

But theoretical research and higher education are not the only strongholds of imperative algorithms. They are also pretty much synonymous with competitive programming, dominating most in-person and online computer science competitions like ICPC and HackerRank, respectively. They are seen as the thing that gifted high school students interested in computer science should pursue – each time said students don’t win medals in the International Olympiad in Informatics or the like, there will be some journalists complaining that their country’s education system is “falling behind”. In many countries, like Poland,² if there’s any high school level computer science education besides basic programming, it will be in imperative algorithms.

Imperative algorithms aren’t just a mere field of study – they are more of a mindset and a culture; following Kuhn’s [5] terminology, they can be said to form a paradigm. Because this paradigm is so immense, so powerful and so entrenched, we feel free to completely disregard it and devote ourselves and this thesis to studying a related field which did not yet reach the status of a paradigm – functional algorithms – focusing on proving their formal correctness.

But before we do that, we spend the rest of this chapter comparing the imperative and functional approaches to algorithms and briefly reviewing available literature on functional algorithms.

¹From now on when we write “algorithms” we will mean “algorithms and data structures”

²I believe the same is true for the rest of the former Eastern Bloc countries too and probably not much better in the West either.

1.2 Two flavours of algorithms

The differences between the fields imperative and functional algorithms are mostly a reflection of the differences between imperative and functional programming languages and only in a small part a matter of differences in research focus.

The basic data structure in imperative languages is the array, which abstracts a contiguous block of memory holding values of a particular type. More advanced data structures are usually records that hold values and pointers/references to other (or even the same) kinds of data structures. The basic control flow primitives for traversing these structures are various loops (`while`, `for`, `do ... while`) and branch/jump statements (`if`, `switch`, `goto`). The most important operation is assignment, which changes the value of a variable (and thus variables do actually vary, like the name suggests [6]). Computation is modeled by a series of operations which change the global state.

In functional languages the basic data structures are created using the mechanism of algebraic data types – elements of each type so defined are trees whose node labels, branching and types of values held are specified by the user. The basic control flow primitives are pattern matching (checking the label and values of the tree's root) and recursion. The most important operation is function composition, which allows building complex functions from simpler ones. Computation is modeled by substitution of arguments for the formal parameters of a function. Variables don't actually vary – they are just names for expressions. [6]

```
int sum(int[] a)
{
    int result = 0;
    for(int i = 0; i < a.length; ++i)
    {
        result += a[i];
    }
    return result;
}
```

Listing 1: A simple program for summing all integers stored in an array, written in an imperative pseudocode that resembles Java.

```

data List a = Nil | Cons a (List a)

sum : List Int -> Int
sum Nil = 0
sum (Cons x xs) = x + sum xs

```

Listing 2: A simple program for summing all integers stored in a (singly-linked) list, written in a functional pseudocode that resembles Haskell.

The two above programs showcase the relevant differences in practice. In both cases we want a function that sums integers stored in the most basic data structure. In the case of our pseudo-Java, this is a built-in array, whereas in our pseudo-Haskell, this is a list defined using the mechanism of algebraic data types, as something which is either empty (`Nil`) or something that has a head of type `a` and a tail, which is another list of values of type `a`.

In the imperative program, we declare a variable `result` to hold the current value of the sum and then we loop over the array. We start by creating an iterator variable `i` and setting it to 0. We successively increment it with each iteration of the loop until it gets bigger than the length of the array and the loop finishes. At each iteration, we modify `result` by adding to it the array entry we’re currently looking at.

In the functional program, we pattern match on the argument of the function. In case it is `Nil` (an empty list), we declare the result to be 0. In case it is `Cons x xs` (a list with head `x` and tail `xs`), we declare that the result is computed by adding `x` and the recursively computed sum of numbers from the list `xs`.

Even though these programs are very simple, they depict the basic differences between imperative and functional algorithms quite well. Some less obvious differences are as follows.

First, functional data structures are by default immutable and thus persistent [7], whereas this is not the case for imperative data structures – they have to be explicitly designed to support persistence. This means implementing some techniques, like backtracking, is very easy in functional languages, but it often requires much more effort in imperative languages. The price of persistence often is, however, increased memory usage.

Second, pointer juggling in imperative languages allows a more efficient implementation of some operations on tree-like structures than using algebraic data types, because nodes in such trees can have pointers not only to their children, but also to parents, siblings, etc. The most famous data structure whose functional, asymptotically optimal implementation is not known is union-find. [8]

The third point is that arrays, the bread-and-butter of imperative programming, provide random access read and write in $O(1)$ time, whereas equivalent functional random access structures work in $O(\log n)$ time where n is the array size (or, at best, $O(\log i)$ where i is the accessed index). This means that algorithms relying on constant time random access will suffer an asymptotic performance penalty when implemented in functional languages. [9]

Even though this asymptotic penalty sounds gloomy, not all hope is lost, because of two reasons. First, mutable arrays can still be used in purely³ functional languages if the mutability is hidden behind a pure interface. An example of this is Haskell’s ST monad. [10] Second, functional languages that are impure, like Standard ML or OCaml, allow using mutable arrays without much hassle, which often saves the day.

1.3 The many-worlds interpretation of the algorithmic business

We stated earlier that we intend to concentrate on formally proving correctness of purely functional algorithms. Before doing that, we shortly peek at how it differs from the activities and workflows of the usual algorithmic business,⁴ what we embrace and what we’re leaving out.

“Necessity is the mother of invention”, goes the proverb, and indeed most algorithms arise as answers to questions like “How do I make a computer sort a list of book titles?” or “How do I arrange my files on the hard disk so that I can easily find them later?”. The path toward the answer usually looks somewhat like this:

- Provide a more precise specification of the problem.
- Design an algorithm that solves the problem and write it down with a good balance of generality and detail.
- Prove that the algorithm is correct. If a proof is hard to find, this may be the sign that the algorithm is incorrect.
- Analyze complexity of the algorithm, i.e. how much resources (number of steps, bits memory, bits of randomness, etc.) does the algorithm need to solve the problem of a given size. If the algorithm needs too much resources, go back and try to design another one that needs less.
- Implement the algorithm and test whether the implementation is correct.

³“Pure” and “impure” are loose terms used to classify functional languages. “Pure” roughly means that a language makes organized effort to separate programs that do ordinary number crunching or data processing from those that have side effects, like throwing exceptions or connecting to a database. An “impure” languages doesn’t attempt at such a separation.

⁴A person engaging in the usual algorithmic business we will call an *algorithmist*.

- Run some experiments to assess the actual performance of the implementation in a few scenarios: random data, data that often occurs in the real world, etc. If the performance is bad, try to find a better implementation or go back and design another algorithm.

Of course this recipe is not stiff and some variations are possible. The most obvious one would be to get rid of the pseudocode and replace it with the actual implementation. A more practice-oriented algorithmist can also skip a difficult analysis in favor of spending more time on performance profiling. A less mathematically gifted algorithmist may skip the proof in favor of more extensive testing etc.

No matter the exact recipe, there is a very enlightening thing to be noticed, namely all the different worlds in which these activities take place. The specification is usually written in English (or whatever the language of the algorithmist), intermingled with some mathematical symbols like quantifiers. The algorithm design process takes place in the algorithmist's mind, but after he's done, it is usually written down in some kind of comfortable pseudocode that allows skipping inessential detail.

The proof is again written in a mixture of English and mathematical symbols, but that's only the surface view. If we take a closer look, it will most likely turn out that the mathematical part is based on classical logic (which admits nonconstructive reasoning principles like proof by contradiction [11] - this can be often seen in proofs of optimality) and some kind of naive set theory, but if pressed a bit, the algorithmist would readily assert that the set theory could be axiomatized using, let's say, the Zermelo-Fraenkel axioms.

The next hidden world, in which the complexity analysis takes its place, is the model of computation. Most likely it won't be mentioned explicitly, just like the logic and set theory thing in the previous paragraph, but we can easily guess it from the fact that the algorithm is intended to be implemented on a real computer and the model of computation most similar to the workings of real hardware is the RAM machine. Last but not least, the implementation, in contrast to the algorithm's pseudocode, will be in some concrete programming language - a very popular one among algorithmists is `C++`.

As we see, the typical solution of an algorithmic problem stems from a sequence (or rather, a loop) of activities which live in quite different worlds. Even though the above recipe and its variations work well in practice, an inquisitive person (perhaps a philosopher) could ask a plethora of questions about how all these worlds fit together:

- Does the pseudocode really describe an algorithm that meets the informally written specification? Is the pen-and-paper proof correct?
- Does the implementation match the pseudocode? Do tests match the specification?

- Does the assumed model of computation match the capabilities of the programming language used for implementation? Is the analysis correct? Do its results match the assessed performance?
-

1.4 An ultra short literature review

Literature review, Okasaki is old and bad for Coq, SF3 is shallow.

1.5 Way of the Coq

Chapter 2

A quick *tour de* sort

Chapter 3

A man, a plan, a canal – MSc thesis

3.1 Design

- First step: specification (describe the path from intuition to formal spec).
- Second step: design (describe the path from the concrete to the abstract by tracing a stub proof of correctness).
- We shouldn't require proofs in order to run programs. Clou: packed vs unpacked classes/records/modules.

3.2 Techniques

- General recursion: Bove-Capretta method as the way to go.
- Functional induction as the way-to-go proof technique. Mention the Equations plugin.

3.3 Topics

- Quicksort: in functional languages we have so powerful abstractions that we can actually implement algorithms and not just programs.
- Braun mergesort: in order not to waste resources, we sometimes have to reify abstract patterns, like the splitting in mergesort.
- Binary heaps: a case study to show the basic workflow and that it's not that obvious how to get basic stuff right.
- Cool data structures: ternary search trees, finger trees.

Bibliography

- [1] <https://www.thefreedictionary.com/algorithm>
- [2] *Do “algorithms” exist in Functional Programming?*,
[https://stackoverflow.com/questions/25940327/
do-algorithms-exist-in-functional-programming](https://stackoverflow.com/questions/25940327/do-algorithms-exist-in-functional-programming)
- [3] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein,
Introduction to Algorithms,
[http://ressources.unisciel.fr/algoprogram/s00aaroot/aa00module1/res/
%5BCormen-AL2011%5DIntroduction_To_Algorithms-A3.pdf](http://ressources.unisciel.fr/algoprogram/s00aaroot/aa00module1/res/%5BCormen-AL2011%5DIntroduction_To_Algorithms-A3.pdf)
- [4] Donald Knuth, *The Art of Computer Programming*
- [5] Thomas S. Kuhn, *The Structure of Scientific Revolutions*
- [6] Robert Harper, *Words Matter*, 2012
<https://existentialtype.wordpress.com/2012/02/01/words-matter/>
- [7] Driscoll JR, Sarnak N, Sleator DD, Tarjan RE
Making data structures persistent, 1986
- [8] Sylvain Conchon, Jean-Christophe Fillâtre,
A Persistent Union-Find Data Structure, 2007
<https://www.lri.fr/~filliatr/ftp/publis/puf-wml07.pdf>
- [9] Nicholas Pippenger, *Pure versus impure Lisp*, 1996
[https://www.cs.princeton.edu/courses/archive/fall03/cs528/
handouts/Pure%20Versus%20Impure%20LISP.pdf](https://www.cs.princeton.edu/courses/archive/fall03/cs528/handouts/Pure%20Versus%20Impure%20LISP.pdf)
- [10] John Launchbury, Simon Peyton Jones,
Lazy Functional State Threads, 1994
[https://www.microsoft.com/en-us/research/wp-content/uploads/1994/
06/lazy-functional-state-threads.pdf](https://www.microsoft.com/en-us/research/wp-content/uploads/1994/06/lazy-functional-state-threads.pdf)
- [11] Andrej Bauer, *Proof of negation and proof by contradiction*, 2010
<http://math.andrej.com/2010/03/29/proof-of-negation-and-proof-by-contradiction/>