FUNCTIONAL PEARL
A fresh look at binary search trees

RALF HINZE
Institute of Information and Computing Sciences, Utrecht University
P.O. Box 80.089, 3508 TB Utrecht, The Netherlands
(e-mail: ralf@cs.uu.nl)

Alle Abstraktion ist anthropomorphes Zerdanken.
— Oswald Spengler, Urfragen

1 Introduction
Binary search trees are old hat, aren’t they? Search trees are routinely covered in introductory computer science classes and they are widely used in functional programming courses to illustrate the benefits of algebraic data types and pattern matching. And indeed, the operation of insertion enjoys a succinct and elegant functional formulation. Fig. 1 contains the six-liner given in the language Haskell 98.

Alas, both succinctness and elegance are lost when it comes to implementing the dual operation of deletion, also shown in Fig. 1. Two additional helper functions are required causing the code size to double in comparison with insertion.

Why this discrepancy? The algorithmic explanation is that insertion always takes place at an external node, that is, at a leaf whereas deletion always takes place at an internal node and that manipulating internal nodes is notoriously more difficult than manipulating external nodes.

Our own stab at explaining this phenomenon is algebraic or, if you like, linguistic. Arguably, the data type Tree with its two constructors, Leaf and Node, does not constitute a particularly elegant algebra. If we use binary search trees for representing sets, then Leaf denotes the empty set \( \emptyset \) and Node \( l \) a \( r \) denotes the set \( s_l \uplus \{ a \} \uplus s_r \) where \( s_l \) and \( s_r \) are the denotations of \( l \) and \( r \), respectively. One might reasonably advance that Node mingles two abstract operations, namely, forming a singleton set \( \{ \cdot \} \) and taking the disjoint union \( \uplus \) of two sets, and that it is preferable to consider these two operations separately.

Of course, there is a good reason for using a ternary constructor: the second argument of Node, the split key, is vital for steering the binary search. Thus, as a replacement for the tree constructors the algebra \( \emptyset, \{ \cdot \}, \uplus \) is inadequate; we additionally need a substitute for the split key. Now, a search tree satisfies the invariant that for each node the split key is greater than the elements in the left subtree (and smaller than the ones in the right subtree). This suggests to augment the algebra with an observer function max (or min, equivalently) that determines the maximum (or the minimum) element of a set. We will see that all standard operations
data Tree a = Leaf | Node (Tree a) a (Tree a)
insert :: (Ord a) ⇒ a → Tree a → Tree a
insert x Leaf = Node Leaf x Leaf
insert x (Node l a r)
  | x < a = Node (insert x l) a r
  | x == a = Node x r
  | x > a = Node l a (insert x r)
delete :: (Ord a) ⇒ a → Tree a → Tree a
delete x Leaf = Leaf
delete x (Node l a r)
  | x < a = Node (delete x l) a r
  | x == a = join l r
  | x > a = Node l a (delete x r)
join :: Tree a → Tree a → Tree a
join Leaf r = r
join (Node l l1 l2) r = Node l m r
    where (l, m) = split-max l1 l2
split-max :: Tree a → a → Tree a → (Tree a, a)
  | a < l1 = (l1, a)
  | a > ra = (Node rl ra rr)
    where r = split-max rl ra rr
member :: (Ord a) ⇒ a → Tree a → Bool
member x Leaf = False
member x (Node l a r)
  | x < a = member x l
  | x == a = True
  | x > a = member x r

Fig. 1. The standard implementation of binary search trees.

on search trees can be conveniently expressed using this extended algebra. This
does not mean, however, that we abandon binary search trees altogether. Rather,
we shall use the algebra as an interface to the concrete representation of this data
structure. This is the point of the pearl: even concrete data types may benefit from
data structural abstraction.

2 An interface to binary search trees

The following signature provides the aforementioned interface to binary search trees.
In fact, it can be seen as a declaration of an abstract data type—the choice of names
and symbols reflects our intention to use trees for representing sets.

data Set a
  ∅  :: (Ord a) ⇒ Set a
  {·} :: (Ord a) ⇒ a → Set a
  (∈) :: (Ord a) ⇒ Set a → Set a → Set a
  max :: (Ord a) ⇒ Set a → a
The constructor $\emptyset$ denotes the empty set, $\{\} \emptyset$ forms a singleton set, and $s_l \uplus s_r$, takes the disjoint union of $s_l$ and $s_r$ under the proviso that the elements in $s_l$ precede the elements in $s_r$. For each constructor there is a corresponding destructor (typeset with a bar) that can be used in patterns: $\emptyset$ matches the empty set, $\{\}$ matches singleton sets, and $s_l \uplus s_r$ matches sets with at least two elements. In the latter case, we may assume that $\max s_l < \min s_r$ and furthermore that both $s_l$ and $s_r$ are non-empty. Thus, the patterns $\emptyset$, $\{a\}$, and $s_l \uplus s_r$ are exhaustive and exclusive. The operation $\max$ is used to determine the maximum element of a non-empty set. We guarantee that all operations, constructors as well as destructors, have a running time that is bounded by a constant.

The signature is asymmetrical in that we provide constant access to the maximum element but not to the minimum element. This will be rectified in Sec. 6. For the moment, we simply note that $\min$ can be defined as a derived operation—albeit with a running time proportional to the height of a tree.

\[
\begin{align*}
\min & \quad \text{(Ord $a$) \Rightarrow Set a \to a} \\
\min \{a\} & = a \\
\min (s_l \uplus s_r) & = \min s_l
\end{align*}
\]

In the second equation we employ the invariants that $\max s_l < \min s_r$ and that $s_l$ is non-empty.

At this point the reader may wonder why it is necessary to distinguish between constructors and destructors? First, the constructors will be implemented by Haskell functions and Haskell does not allow functions to appear in patterns. Second, the expression $\emptyset \uplus \{a\}$ must not be equal to the pattern $\emptyset \uplus \{a\}$ since we wish to guarantee that both arguments of the destructor `\uplus` are non-empty. In fact, we have $s_l \uplus s_r = s_l \uplus s_r$ if and only if both $s_l$ and $s_r$ are non-empty. On the other hand, $\emptyset = \emptyset$ and $\{a\} = \{a\}$ hold unconditionally.

## 3 Set functions

Given the above interface we can easily define the standard operations on sets.

Here is how we implement set membership.

\[
\begin{align*}
\text{member} & \quad \text{(Ord $a$) \Rightarrow a \to Set a \to \text{Bool}} \\
\text{member } x \emptyset & = \text{False} \\
\text{member } x \{a\} & = x = a \\
\text{member } x (s_l \uplus s_r) & \\
| \quad x \leq \max s_l & = \text{member } x s_l \\
| \quad \text{otherwise} & = \text{member } x s_r
\end{align*}
\]

The recursive structure of the definition is archetypical and nicely illustrates a separation of concerns. Elements of a set are accessed solely through the pattern $\{a\}$ whereas the pattern $s_l \uplus s_r$ in conjunction with the observer function `\max` is used for implementing the divide-and-conquer step. In other words, the operations on elements always take place at the fringe of the tree.
Insertion and deletion are now equally simple to implement.

\[
\begin{align*}
\text{insert} & :: \ (\text{Ord} \ a) \Rightarrow a \to \text{Set} \ a \to \text{Set} \\
\text{insert} \ x \ \emptyset & = \{x\} \\
\text{insert} \ x \ \{a\} & =
\begin{cases}
  \{x\} \cup \{a\}, & x < a \\
  \{x\}, & x = a \\
  \{a\} \cup \{x\}, & x > a
\end{cases} \\
\text{insert} \ x \ (s_l \uplus s_r) & =
\begin{cases}
  \text{insert} \ x \ s_l \uplus s_r, & x \leq \text{max} \ s_l \\
  s_l \uplus \text{insert} \ x \ s_r, & \text{otherwise}
\end{cases}
\end{align*}
\]

\[
\begin{align*}
\text{delete} & :: \ (\text{Ord} \ a) \Rightarrow a \to \text{Set} \ a \to \text{Set} \\
\text{delete} \ x \ \emptyset & = \emptyset \\
\text{delete} \ x \ \{a\} & =
\begin{cases}
  \emptyset, & x = a \\
  \{a\}, & \text{otherwise}
\end{cases} \\
\text{delete} \ x \ (s_l \uplus s_r) & =
\begin{cases}
  \text{delete} \ x \ s_l \uplus s_r, & x \leq \text{max} \ s_l \\
  s_l \uplus \text{delete} \ x \ s_r, & \text{otherwise}
\end{cases}
\end{align*}
\]

Note that the two functions differ in the treatment of the base cases only.

The definition of \text{delete} can be slightly simplified for the special case that we remove the maximum element.

\[
\begin{align*}
\text{delete-max} & :: \ (\text{Ord} \ a) \Rightarrow \text{Set} \ a \to \text{Set} \\
\text{delete-max} \ \emptyset & = \emptyset \\
\text{delete-max} \ \{a\} & = \emptyset \\
\text{delete-max} \ (s_l \uplus s_r) & = s_l \uplus \text{delete-max} \ s_r
\end{align*}
\]

The functions \text{max} and \text{delete-max} provide priority queue functionality—except that priority queues are usually bags rather than sets.

### 4 Implementing the interface

Recall from Sec. 2 that we have to guarantee that none of the operations takes more than a constant number of steps. Clearly, this condition rules out `standard` binary search trees as the underlying data structure: determining the maximum element in a search tree takes time proportional to the length of the right spine. However, this observation suggests that we might meet the desired time bound if we constrain the length of the right spine. We take the simplest approach and restrict ourselves to search trees where the right subtree of the root is empty. The following data declaration makes this restriction explicit.

\[
\text{data} \ Set \ a = \text{Leaf} \mid \text{Root} \ (Set \ a \ a) \mid \text{Node} \ (Set \ a) \ a \ (Set \ a)
\]

The term \text{Root} \ t \ a serves as a replacement for the top-level term \text{Node} \ t \ a \ \text{Leaf}. Thus, a search tree is either empty or of the form \text{Root} \ t \ a—we insist that \text{Root} is used only at the top level and that \text{Node} only appears below a \text{Root} constructor.
Given this representation it is straightforward to implement the constructors $\emptyset$, \{\}, `\$` and the observer function $\textbf{max}$.

\[
\emptyset = \text{Leaf} \\
\{a\} = \text{Root Leaf } a \\
\text{Leaf} \oplus t_r = t_r \\
t_l \oplus \text{Leaf} = t_l \\
\text{Root } t_l a_l \oplus \text{Root } t_r a_r = \text{Root } (\text{Node } t_l a_l t_r) a_r \\
\textbf{max } (\text{Root } t a) = a
\]

To implement the destructors $\emptyset$, \{\}, and `\$` we make use of an extension to Haskell 98 called views (Burton \textit{et al.}, 1996; Okasaki, 1998). Briefly, a view allows any type to be viewed as a free data type. A view declaration for a type $T$ consists of an anonymous data type, the view type, and an anonymous function, the view transformation, that shows how to map elements of $T$ to the view type.

\[
\textbf{view} \text{ Set } a = \emptyset | \{a\} | \text{Set } a \oplus \text{Set } a \text{ where} \\
\text{Leaf} \rightarrow \emptyset \\
\text{Root Leaf } a \rightarrow \{a\} \\
\text{Root } (\text{Node } t_l a_l t_r) a_r \rightarrow \text{Root } t_l a_l \oplus \text{Root } t_r a_r
\]

The view transformation essentially undoes the work of the constructors—it is not the inverse since, for instance, $\emptyset \oplus \{a\}$ matches $\{a\}$ rather than $s_l \oplus s_r$.

\section{Eliminating the abstraction layer}

Worried about efficiency? It is a simple exercise in program fusion to eliminate the anonymous view type from the definitions given in Sec. 3—a good optimizing compiler should be able to perform this transformation automatically. However, since the resulting code is instructive from an algorithmic point of view, let us briefly discuss one example.

In general, each of the set functions can be written as a composition of the view transformation and the `original` function that works on the view type. Since the view type is non-recursive, we can easily fuse the view transformation and the original function. In the case of set membership we obtain the following definition.

\[
\text{member } x \text{ Leaf} = \text{False} \\
\text{member } x \text{ (Root Leaf } a) = x \equiv a \\
\text{member } x \text{ (Root } (\text{Node } t_l a_l t_r) a_r) \\
\begin{align*}
| x \leq a_l & = \text{member } x \text{ (Root } t_l a_l) \\
| \text{otherwise} & = \text{member } x \text{ (Root } t_r a_r)
\end{align*}
\]

Note that in both recursive calls \textit{member} is passed a \textit{Root} node that is constructed on the fly. As a simple optimization we avoid building this intermediate term by
specializing member $x (\text{Root } t \ a)$ to member$^t x \ t \ a$.

\[
\begin{align*}
\text{member } x & \text{ Leaf} = \text{False} \\
\text{member } x (\text{Root } t \ a) & = \text{member}^t x \ t \ a \\
\text{member}^t x & \text{ Leaf } a = x \triangleleft a \\
\text{member}^t x (\text{Node } t_l \ a_l \ t_r \ a_r) & = \begin{cases} \\
   \text{member}^t x \ t_l \ a_l & \text{if } x \leq a_l \\
   \text{member}^t x \ t_r \ a_r & \text{otherwise}
\end{cases}
\end{align*}
\]

Interestingly, this implementation of member closely resembles an algorithm proposed by Andersson (1991). Recall that the standard implementation of set membership shown in Fig. 1 uses one three-way comparison per visited node. The variant of Andersson manages with one two-way comparison by keeping track of a candidate element that might be equal to the query element. The third argument of member$^t$ exactly corresponds to this candidate element, which is only checked for equality when a leaf is hit.

6 A more symmetric design

The implementation in Sec. 4 supports a constant time max operation but not a constant time min operation. In this section we show how to symmetrize the implementation so that both operations can be supported in constant time. This time we deviate slightly from binary search trees.

Currently, the Root constructor only contains the maximum element of the represented set. An obvious idea is to add a third field to the constructor which contains the minimum. This, however, implies that we can no longer represent singleton sets—unless we are willing to allow both fields to contain the same element. Instead, we introduce a new unary constructor that forms a singleton.

Of course, we have to make sure that we can still take the disjoint union of two sets in constant time. This is easily done if one of the sets is empty or both are singletons. In the latter case, we construct a new Root node with an empty subtree. Now, assume that both arguments are of the form Root $a_l \ t \ a_r$. In this case, we have to form an internal tree using two subtrees and two split keys. Similarly, if one of the arguments is a singleton and the other one has Root as the topmost constructor, then we have to build a tree using one subtree and one split key. For each of the three cases, we invent a tailor-made constructor:

\begin{verbatim}
data Set a = Leaf | Single a | Root a (Set a) a | Cons a (Set a) | Snoc (Set a) a | Node (Set a) a (Set a).
\end{verbatim}

We insist that Single and Root are only used at the top level and that Cons, Snoc, and Node only appear below a Root node. Given this data structure the implementation of the interface is straightforward (Fig. 2 lists the code).
Fig. 2. An implementation supporting constant time \textit{min}- and \textit{max}-operations.

This variation nicely illustrates the merits of abstraction. Since the set functions of Sec. 3 only rely on the abstract interface, they happily work with the new implementation. Another interesting variation is to augment the implementation of Sec. 4 by a balancing scheme. An extension along this line is described in Hinze (2001) albeit for the more elaborate data structure of priority search queues. All in all, a refreshing view on an old data structure.

References


