Optimal segmentations

Optimal segmentations

by

J.S.C.P. van der Woude

89/15

December, 1989

COMPUTING SCIENCE NOTES

This is a series of notes of the Computing Science Section of the Department of Mathematics and Computing Science Eindhoven University of Technology.
Since many of these notes are preliminary versions or may be published elsewhere, they have a limited distribution only and are not for review.
Copies of these notes are available from the author or the editor.

Eindhoven University of Technology
Department of Mathematics and Computing Science
P.O. Box 513
5600 MB EINDHOVEN
The Netherlands

Editors: prof.dr.M.Rem
prof.dr.K.M. van Hee

OPTIMAL SEGMENTATIONS

Introduction

In programming methodology the attention gradually shifts from specific problems towards classes of problems, their characterization and theorems for their solutions. A classification of segment problems is in progress and several solution schemes may be viewed as theorems. A type of problems not too distant from the segment problems is that of partitionings. Given a sequence (or set) construct a partition, possibly an extremal partition, whose members all satisfy certain conditions. E.g. partition a list into segments that satisfy a certain "nice" predicate, give a construction of a partition with as few members as possible; such a partition may be called an optimal segmentation. I'll derive conditions on the predicate involved that guarantee efficient algorithms modulo the predicate calculations (i.e. evaluation of predicates is assumed to take constant time). Moreover, it is shown that the proposed algorithms are greedy.

Notation and concepts

One of the alleged disadvantages of predicate calculus notation is indexitis. This is often circumvented by introduction of abbreviations and ad hoc notations. A more compact, sometimes even too compact, notation is the so-called Bird-Meertens formalism (with APL rudiments, see [B]). Just as an experiment, I incorporate some of the BM features in predicate notation.

For a set (type) α, the triple $(\alpha*, +\!\!+, [\,])$ denotes the monoid of lists over α.
Lists are denoted as sequences between brackets. The catenation $(+\!\!+)$[1] and the unit ($[\,]$, the empty list) are polymorphic. So lists ($\alpha*$) as well as lists of lists ($\alpha**$) are both considered with the same symbols for catenation and unit, the distinction may be seen from the choice of identifiers:

$$a \in \alpha$$

$$u, v, \ldots, z \in \alpha*$$

$$us, vs, \ldots, zs \in \alpha**$$

I'll use reduction (just $+\!\!+/$, flatten) and filter ($\lhd$) as in BM. The functions inits, tails and segs are considered in the set-valued versions of those in BM, e.g.:

$$\text{tails}.xs = \{vs \mid (\mathbf{E}us \;::\; xs = us +\!\!+ vs)\} \;.$$

The segmentation concepts are formalized as follows:

[1] As an experiment, $+\!\!+$ will be given the highest priority: $f.x +\!\!+ y = f.(x +\!\!+ y)$.

Let $Q : \alpha* \to \text{Bool}$ be a predicate on α-lists.
Define the relations $\mathcal{P}, \mathcal{OP} \subseteq \alpha** \times \alpha*$ and
the function $N : \alpha* \to I\!N$ by

- $xs\mathcal{P}x \equiv +\!\!+/xs = x \wedge Q \lhd xs = xs$
- $N.x = (\downarrow xs : xs\mathcal{P}x : \#xs)$
- $xs\mathcal{OP}x \equiv xs\mathcal{P}x \wedge N.x = \#xs$

Then $xs(\mathcal{O})\mathcal{P}x$ may be paraphrazed as: xs is an (optimal) Q-segmentation for x.
Note that optimal Q-segmentations need not be unique.

Some properties

It is good practice to collect, prior to the derivation, some properties of the concepts involved. The easy proofs are left as exercises:

(0) $[]\mathcal{P}[]$, hence $N.[] = 0$ and $[]\mathcal{OP}[]$

(1) $xs\mathcal{P}x \wedge ys\mathcal{P}y \Rightarrow xs +\!\!+ ys\,\mathcal{P}x +\!\!+ y$

(2) $xs\mathcal{P}x \wedge us \in \text{segs}.xs \Rightarrow us\mathcal{P} +\!\!+/us$

(3) $xs\mathcal{OP}x \wedge us \in \text{segs}.xs \Rightarrow us\mathcal{OP} +\!\!+/us$

(4) $xs +\!\!+ [[]] +\!\!+ ys\,\mathcal{P}x \Rightarrow xs +\!\!+ ys\mathcal{P}x$

(5) Note that by (4), empty segments may be discarded in considering optimal segmentations. If necessary one may consider Q' with $Q'.x \equiv Q.x \wedge x \neq []$ instead of Q.

Life would have been a lot easier (although very dull) if the $\mathcal{OP}$ version of (1) were true, quod non. Since the $\mathcal{P}$-part of $\mathcal{OP}$ behaves nicely, an investigation of N is in order. It seems interesting to see whether some recurrence is lurking around. Indeed

(6) $N.x +\!\!+ [a] = (\downarrow z, w : w +\!\!+ z = x \wedge Q.z +\!\!+ [a] : N.w + 1)$

For: $N.x +\!\!+ [a]$

$= \{\text{def } N\}$

$(\downarrow ys : ys\mathcal{P}x +\!\!+ [a] : \#ys)$

$= \{+\!\!+/ys = x +\!\!+ [a] \Rightarrow ys \neq []\}$

$(\downarrow zs, z : zs +\!\!+ [z]\mathcal{P}x +\!\!+ [a] : \#zs + 1)$

$= \{\text{def } \mathcal{P}\}$

$(\downarrow zs, z : (+\!\!+/zs) +\!\!+ z = x +\!\!+ [a] \wedge Q \lhd zs = zs \wedge Q.z : \#zs + 1)$

$= \{\text{one point rule}\}$

$(\underline{\downarrow} zs, z, w \ : \ w{+\!\!+}z = x{+\!\!+}[a] \ \wedge \ w = {+\!\!+}/zs \ \wedge \ Q \lhd zs = zs \ \wedge \ Q.z \ : \ \#zs + 1)$

$= \{\text{def } \mathcal{P}\}$

$(\underline{\downarrow} zs, z, w \ : \ w{+\!\!+}z = x{+\!\!+}[a] \ \wedge \ zs \mathcal{P} w \ \wedge \ Q.z \ : \ \#zs + 1)$

$= \{\text{promotion}\}$

$(\underline{\downarrow} z, w \ : \ w{+\!\!+}z = x{+\!\!+}[a] \ \wedge \ Q.z \ : \ (\underline{\downarrow} zs \ : \ zs \mathcal{P} w \ : \ \#zs + 1))$

$= \{\text{def } N, \ \text{pinf} + 1 = \text{pinf}\}$

$(\underline{\downarrow} z, w \ : \ w{+\!\!+}z = x{+\!\!+}[a] \ \wedge \ Q.z \ : \ N.w + 1)$

$= \{\text{split off } z = [\,], \text{ without loss of generality } \neg Q.[\,] \ \ (5)\}$

$(\underline{\downarrow} z, w \ : \ w{+\!\!+}z = x \ \wedge \ Q.z{+\!\!+}[a] \ : \ N.w + 1)$

Note that, thanks to the rule pinf + 1 = pinf, the validity of the recurrence relation is independent of the existence of Q-segmentations. Nonexistence is rather unsatisfactory, so I propose an easy way out: assume

(7) $Q.[a]$ for every $a \in \alpha$

Hence the exotic rule pinf + 1 = pinf is superfluous.

Thinning out the quantification

Since in the recurrence relation a quantification over all postfixes of x occurs, the resulting algorithm is quadratic modulo Q-calculations. Efficiency improvement is to be expected if only a small subset of the postfixes of x suffices. Given an optimal Q-segmentation xs for x an interesting subset of the postfixes of x is given by

$$\{{+\!\!+}/vs \mid vs \in \text{tails}.xs\} \ \ (=: \ T) \ .$$

In order to restrict the quantification in the right-hand side of (6) to $z \in T$, there should be reasons to discard $z \notin T$. Consider the following Setting (S)

(S)
- (i) $x = {+\!\!+}/xs \ \wedge \ x = w{+\!\!+}z \ \wedge \ z \notin T$
- (ii) $xs \mathcal{O}\mathcal{P} x \ \wedge \ Q.z{+\!\!+}[a]$

By (i), there are us, vs, u, v such that

$xs = us{+\!\!+}[u{+\!\!+}v]{+\!\!+}vs$ and

$w = ({+\!\!+}/us){+\!\!+}u \ \wedge \ z = v{+\!\!+}({+\!\!+}/vs) \ \wedge \ u \neq [\,] \ \wedge \ v \neq [\,] \ .$

One may forget about this z in the quantification of (6) if there is a Q-segmentation zs of $x +\!\!+ [a]$ such that

- $\text{last}.zs = p +\!\!+ [a]$ for some $p \in T$
- $\#zs \leq N.w + 1$

Given setting (S), two obvious candidates for zs can be constructed from the Q-segmentation xs, such that $\text{last}.zs = p +\!\!+ [a]$ for some $p \in T$:

(c0) $zs = us +\!\!+ [u +\!\!+ v +\!\!+ (+\!\!+/vs) +\!\!+ [a]]$

(c1) $zs = us +\!\!+ [u +\!\!+ v] +\!\!+ [(+\!\!+/vs) +\!\!+ [a]]$

These candidates are Q-segmentations if:

- ad (c0): $Q.u +\!\!+ v +\!\!+ (+\!\!+/vs) +\!\!+ [a]$
  Since $u +\!\!+ v$ in xs and $xsPx$, certainly $Q.u +\!\!+ v$.
  By (ii), $Q.z +\!\!+ [a]$, while $z = v +\!\!+ (+\!\!+/vs)$ and $v \neq []$ ((S)).
  Hence overlap closedness of Q is sufficient.
  (I.e. $Q.k +\!\!+ l \ \wedge \ Q.l +\!\!+ m \ \wedge \ l \neq [] \ \Rightarrow \ Q.k +\!\!+ l +\!\!+ m.$)

- ad (c1): $Q.(+\!\!+/vs) +\!\!+ [a]$
  Since $Q.z +\!\!+ [a]$, while $z = v +\!\!+ (+\!\!+/vs)$, it is sufficient to require Q to be postfix closed.
  (I.e. $Q.k +\!\!+ l \Rightarrow Q.l$. Indeed a weaker requirement could be
  $Q.k +\!\!+ l \ \wedge \ Q.l +\!\!+ m \ \wedge \ l \neq [] \ \Rightarrow \ Q.m$,
  which seems a somewhat awkward property.)

With respect to the last requirement:

$$
\begin{array}{ll}
 & \#zs \leq N.w + 1 \\
= & \{\#zs = \#us + 1 + j \text{ for candidate (cj)}\} \\
 & \#us \leq N.w - j \\
\Leftarrow & \{\text{In setting (S)} : \ us \sqsubset xs \ \wedge \ +\!\!+/us \sqsubset w \sqsubset +\!\!+/xs\} \\
\text{(OSj)} & (\mathbf{A}us', w' \ : \ us' \sqsubset xs \ \wedge \ +\!\!+/us' \sqsubset w' \sqsubset +\!\!+/xs \ : \ \#us' \leq N.w' - j)
\end{array}
$$

where "$\sqsubseteq$" denotes the prefix order:

$$p \sqsubseteq q \equiv p \in \text{inits}.q \ , \quad p \sqsubset q \equiv p \sqsubseteq q \ \wedge \ p \neq q \ .$$

The universal quantification in (OSj) is chosen because

- us and w in the setting (S) are arbitrarily chosen such that $z \notin T$. It is desirable to have a condition that is independent of that choice.
- (OSj) is a property of the Q-segmentation xs alone (even optimality is not used).

The established "thinning out" may be formulated as:

(8) Lemma. Let $xs\,OP\,x$. In each of the following two cases:

$L0$: Q is overlap closed and xs satisfies OS0

$L1$: Q is postfix closed and xs satisfies OS1

the quantification in (6) may be thinned out to

$$N.x +\!\!+ [a] = (\underline{\downarrow} us, vs \;:\; us +\!\!+ vs = xs \;\wedge\; Q.(+\!\!+/vs) +\!\!+ [a] \;:\; \#us + 1)\,.$$

Proof.

$$\begin{aligned}
& N.x +\!\!+ [a] \\
= \;& \{(6),\ Lj \text{ hence restriction to } z \in T\} \\
& (\underline{\downarrow} w, z \;:\; z \in T \;\wedge\; w +\!\!+ z = x \;\wedge\; Q.z +\!\!+ [a] \;:\; N.w + 1) \\
= \;& \{z \in T \equiv (\mathbf{E} us, vs \;:\; us +\!\!+ vs = xs \;:\; z = +\!\!+/vs)\,; \text{calc}\} \\
& (\underline{\downarrow} us, vs \;:\; us +\!\!+ vs = xs \;: \\
& \qquad (\underline{\downarrow} w, z \;:\; z = +\!\!+/vs \;\wedge\; w +\!\!+ z = x \;\wedge\; Q.z +\!\!+ [a] \;:\; N.w + 1)) \\
= \;& \{+\!\!+/xs = x \text{ and } w +\!\!+ z = (+\!\!+ (us/ +\!\!+ z \equiv w = +\!\!+/us\} \\
& (\underline{\downarrow} us, vs \;:\; us +\!\!+ vs = xs \;\wedge\; Q.(+\!\!+/vs) +\!\!+ [a] \;:\; N.(+\!\!+/us) + 1) \\
= \;& \{xs\,\mathcal{OP}\,x \;\wedge\; us \sqsubseteq xs,\ (3)\} \\
& (\underline{\downarrow} us, vs \;:\; us +\!\!+ vs = xs \;\wedge\; Q.(+\!\!+/vs) +\!\!+ [a] \;:\# us + 1)
\end{aligned}$$

□

Lemma (8) only guarantees efficiency improvement if the (OSj) property is an invariant in the (successive) construction of optimal segmentations. This will be addressed in the next section.

Construction of an optimal segmentation

In the following blueprint for the calculation of an optimal segmentation for $X \in \alpha*$, only the invariance of $I2$ is left to be proved:

$I0$ $\quad x +\!\!+ x' = X$

$I1$ $\quad xs\,\mathcal{OP}\,x$

$I2$ $\quad xs$ satisfies OSj

$$
\begin{array}{l}
\quad x\,,x',xs \;:=\; [\,]\,,X\;,\;[\,]\;\;\{I\}\\
;\,\textbf{do}\;\; x' \neq [\,]\\
\quad\longrightarrow\; a \;:=\; \mathrm{hd}.x'\\
\qquad ;\, S\;\; \{(ys,zs) \text{ is a witness for}\\
\qquad\qquad (\underline{\downarrow}(us,vs)\;:\; us{+\!\!+}vs = xs \;\wedge\; Q.(+\!\!+/vs){+\!\!+}[a]\;:\; \#us+1)\}\\
\qquad ;\, xs \;:=\; ys{+\!\!+}[(+\!\!+/zs){+\!\!+}[a]]\;\;\{I1[x \;:=\; x{+\!\!+}[a]] \wedge I2!\}\\
\qquad ;\, x,x' \;:=\; x{+\!\!+}[a],\mathrm{tl}.x'\;\;\{I\}\\
\quad \textbf{od}\;\;\{I \;\wedge\; x = X, \text{ hence } xs\mathcal{OP}X\}
\end{array}
$$

In order to prove the invariance of $I2$, assume

(i) $+\!\!+/(ys{+\!\!+}[q]) = (+\!\!+/xs){+\!\!+}[a]$ $\quad\{\text{where } q = (+\!\!+/zs){+\!\!+}[a]\}$

(ii) $ys \sqsubseteq xs$ $\quad\{(ys,zs) \text{ is a witness}\}$

(iii) $N.+\!\!+/xs = \#xs$ $\quad\{I1 \;\wedge\; \mathrm{def}.N\}$

then
$$
\begin{array}{l}
\quad ys{+\!\!+}[q] \text{ satisfies OSj}\\
=\;\{\text{def OSj}\}\\
\quad (\mathbf{A}us,w\;:\; us \sqsubset ys{+\!\!+}[q] \;\wedge\; +\!\!+/us \sqsubset w \sqsubset +\!\!+/(ys{+\!\!+}[q])\;:\; \#us \leq N.w - j)\\
=\;\{(\mathrm{i})\,;\sqsubseteq\}\\
\quad (\mathbf{A}us,w\;:\; us \sqsubseteq ys \;\wedge\; +\!\!+/us \sqsubset w \sqsubseteq +\!\!+/xs\;:\; \#us \leq N.w - j)\\
\Leftarrow\;\{((\mathrm{ii})\,;\text{split off } w = +\!\!+/xs\;;\; +\!\!+/us \sqsubset +\!\!+/xs \;\Rightarrow\; us \sqsubset xs\}\\
\quad (\mathbf{A}us,w\;:\; us \sqsubset xs \;\wedge\; +\!\!+/us \sqsubset w \sqsubset +\!\!+/xs\;:\; \#us \leq N.w - j)\\
\qquad\quad \wedge\; (\mathbf{A}us\;:\; us \sqsubset xs\;:\; \#us \leq N.+\!\!+/xs - j)\\
=\;\{\text{def OSj}\,;(\mathrm{iii}) \text{ and } j \in \{0,1\}\}\\
\quad xs \text{ satisfies OSj } (\wedge \text{ true})
\end{array}
$$

Note that OS1 is an invariant for the construction in both cases, Q is overlap closed and Q is postfix closed.

For the construction of S in case Q is overlap closed I don't see a better solution than just checking all splittings of xs. However, in case Q is postfix closed, things are a lot more attractive: since

$$\neg Q.q \;\Rightarrow\; \neg Q.p{+\!\!+}q$$

S boils down to a linear search:

$$
\begin{array}{l}
ys, zs, q \;:=\; xs, [\,], [a] \\
\{ys +\!\!+ zs = xs \;\wedge\; Q.q \;\wedge\; q = (+\!\!+/zs) +\!\!+ [a]\} \\
\;; \textbf{do} \quad ys \neq [\,] \;\underline{\text{cand}}\; Q.(\text{last}.ys) +\!\!+ q \\
\quad \longrightarrow \; ys, zs, q \;:=\; \text{front}.ys \;,\; [\text{last}.ys] +\!\!+ zs \;,\; (\text{last}.ys) +\!\!+ q \\
\quad \textbf{od}
\end{array}
$$

S can easily be mixed with the assignment to xs. [Identify ys and xs, forget about zs in the above].
The complete algorithm is linear (modulo Q-calculations) which is evident from the variant function

$$2 * \#X - 2 * \#x + \#xs \;.$$

For completeness sake: the algorithm, in case Q is postfix closed, is:

$$
\begin{array}{l}
x, x', xs \;:=\; [\,], X, [\,] \\
\;; \textbf{do} \quad x' \neq [\,] \\
\quad \longrightarrow \; a \;:=\; \text{hd}.x' \;;\; q \;:=\; [a] \\
\qquad ; \textbf{do} \quad xs \neq [\,] \;\underline{\text{cand}}\; Q.(\text{last}.xs) +\!\!+ q \\
\qquad\quad \longrightarrow \; xs, q \;:=\; \text{front}.xs \;,\; (\text{last}.xs) +\!\!+ q \\
\qquad\quad \textbf{od} \\
\qquad ; x, x', xs \;:=\; x +\!\!+ [a], \text{tl}.x', xs +\!\!+ [q] \\
\quad \textbf{od}
\end{array}
$$

Greedy Q-segmentations

Interpretation of the strongest OS condition (OS1) leads to some feeling of greediness. The definition of (left-) greediness for Q-segmentations (see [B]):

(9) $\quad \text{Greedy}.[\,]$

$$\text{Greedy}.[x] +\!\!+ xs \;\equiv\; \text{Greedy}.xs \;\wedge\; x = (\underline{\uparrow} z \;:\; z \sqsubseteq x +\!\!+ (+\!\!+/xs) \;\wedge\; Q.z \;:\; z)$$

The following lemma shows that the construction in the former section is a construction for the greedy Q-segmentation:

(10) Lemma. Let xs be a Q-segmentation with $Q.{+\!\!+}/xs \equiv \#xs \leq 1$. Then

$$xs \text{ satisfies OS1} \Rightarrow \text{Greedy}.xs.$$

Proof. By induction on $\#xs$. The base-case, $\#xs \leq 1$, is trivial.
Suppose $\#xs \geq 1$. Then for Q-segmentation $[x]{+\!\!+}xs$:

$[x]{+\!\!+}xs$ satisfies OS1

$\Rightarrow$ {domain restriction}

$(\mathbf{A}us, w : [x] \sqsubseteq us \sqsubset [x]{+\!\!+}xs \wedge {+\!\!+}/us \sqsubset w \sqsubset x{+\!\!+}({+\!\!+}/xs) : \#us < N.w)$

$=$ {dummy change for us, w}

$(\mathbf{A}us, w : us \sqsubset xs \wedge {+\!\!+}/us \sqsubset w \sqsubset {+\!\!+}/xs : \#us + 1 < N.x{+\!\!+}w)$

$\Rightarrow$ {$Q.x$, so $N.x{+\!\!+}w \leq 1 + N.w$; def OS1}

xs satisfies OS1

$\Rightarrow$ {Ind. hyp.}

Greedy.xs

and

$[x]{+\!\!+}xs$ satisfies OS1

$\Rightarrow$ {instantiate $us := [x]$; $\#xs \geq 1$}

$(\mathbf{A}w : x \sqsubset w \sqsubset x{+\!\!+}({+\!\!+}/xs) : 1 < N.w)$

$\Rightarrow$ {$1 < N.w \Rightarrow 1 \neq N.w$; $w \neq [\,] \Rightarrow (1 = N.w \equiv Q.w)$}

$(\mathbf{A}w : x \sqsubset w \sqsubset x{+\!\!+}({+\!\!+}/xs) : \neg Q.w)$

$=$ {$\#([x]{+\!\!+}xs) > 1 \Rightarrow \neg Q.x{+\!\!+}({+\!\!+}/xs)$; $Q.x$}

$x = (\uparrow w : w \sqsubseteq x{+\!\!+}({+\!\!+}/xs) \wedge Q.w : w)$ □

Afterthought and acknowledgements

The derivation of the requirements on Q and the corresponding algorithms were what I was after. However, also the solutions themselves are interesting: The shape of the "postfix-closed" version is very familiar. It has a striking resemblance with the algorithms for

- the maximal pre- and postfix of a string [(W0)].
- the largest rectangle under a histogram [(W1)].
- the maximal one-sided extreme segment (and the like).

A common root for all these problems would be very interesting. I don't mean simply the use of a stack that is apparent in these examples, but a general recognition strategy and a theorem that converts the recognition (almost) immediately into an algorithm.

The problem and the challenge to derive the solution resulted from discussions in the algorithmics working group at the Rijks Universiteit van Utrecht. Hans Zantema gave a functional solution using a direct proof that greedy is optimal. The solution presented here inspired Maarten Fokkinga to give a full account of promotion possibilities for an optimal segmentation problem, leading to a kind of "taxonomy" of their solution schemes ([F]). Oege de Moor presented a Bird-Meertens derivation in Ameland ([M]).

References

[B] Bird, R.S., An introduction to the theory of lists, in NATO ASI, Series F, vol 36, Springer (1987).

[F] Fokkinga, M., Squiggolish derivations for ....., Lecture Notes (part III), Hollum-Ameland (1989).

[M] Moor, O. de, List partitions, Lecture Notes (part II), Hollum-Ameland (1989).

[W0] Woude, J.C.S.P. van der, Playing with patterns searching for strings, SCP ..... . *

[W1] Woude, J.C.S.P. van der, Rabbitcount := Rabbitcount-1, in "Groningen 375". **

* SCP 12 (1989) 99-118

** LNCS 375 (1989).