Franz Stetter

Estimates in software science
ESTIMATES IN SOFTWARE SCIENCE

FRANZ STETTER

Abstract: A program length estimate which depends on the vocabulary size only is presented and it is shown that this estimate should substitute the formula based on the count of the operators and operands. Then, the length estimate obtained is used to establish a simple relationship between the vocabulary size and the potential vocabulary size. Furthermore, the length estimate obtained can be adjusted such that it is exact at its lower bound, namely for the case the vocabulary size complies with the potential vocabulary size.

Index Terms: Complexity measures, program estimates, program predictions, software metrics, software science.
I. Length estimate

Software science [1] deals with measurable properties of algorithms written in a programming language. Basic variables are:

- $n$: vocabulary size,
- $n_1$: unique operator count,
- $n_2$: unique operand count,
- $n^*$: potential vocabulary size,
- $\lambda$: language level,
- $N$: program length,
- $N_1$: total number of operators,
- $N_2$: total number of operands,
- $V$: program volume,
- $V^*$: potential program volume,
- $V^*_b$: boundary program volume.

Frequently used equations are

\[ n = n_1 + n_2, \quad N = N_1 + N_2, \quad (1) \]

and the estimate $\hat{N}_H$ [1] for $N$:

\[ \hat{N}_H = n_1 \log n_1 + n_2 \log n_2 \quad (2) \]

In this paper, we show that the estimate

\[ \hat{N} = n \log (n/2) = n(\log n - 1) \quad (3) \]

is as accurate as (2). Since $\hat{N}$ depends on one variable only, equation (3) is preferable to equation (2). Moreover, the quantities $n_1, n_2, N_1, N_2$ are used in software science by means of the equations (1) in almost all cases but, by substituting (2) by (3), we do not have to distinguish between operators and operands. It is sufficient to take the size of the vocabulary and of the potential vocabulary as basic measures. From there, we can say that a program is composed of "units" which are elements of the vocabulary (a "unit" is the collective name for operators and operands).

In [1], equation (3) is used as an approximation for (2) in cases the term of $\hat{N}_H$ is too complicated to deal with but (3) is not proposed as length estimate actually.
II. Testing \( \hat{N} \)

In [1] and in numerous papers, it was shown that (2) is a very good estimate of \( N \) for well-written programs. Thus, we first have to show that (3) is very close to (2).

Let be

\[
\begin{align*}
\hat{n}_1 &= p \ n, \\
\hat{n}_2 &= (1-p) \ n,
\end{align*}
\]

where \( 0 < p < 1 \), then we get from (2):

\[
\begin{align*}
\hat{N}_H &= p \ n \ln(pn) + (1-p) \ n \ln((1-p)n) \\
&= n \ln(n/2) + n (p \ln p + ((1-p) \ln (1-p) + 1) \\
&= \hat{N} + n f(p).
\end{align*}
\]

The function \( f(p) \) is symmetric to \( p = 1/2 \), decreases monotonically from 0 to 1/2, and it is

\[
0 < f(p) < 1 \quad \text{for} \quad p \in (0, 1).
\]

Thus, the relative error

\[
|\hat{N}_H - \hat{N}| / \hat{N} = |f(p)/\ln(n/2)| < 1/(\ln n - 1)
\]

tends to zero for large \( n \). For smaller values of \( n \), we compare the estimates \( \hat{N}_H \) and \( \hat{N} \) using the examples of table 5.1 of [1] - among them are algorithms (1) to (14) of Comm. of ACM. The following table demonstrates a very close correspondence between \( \hat{N} \) and \( N \) which is even slightly better than that between \( \hat{N}_H \) and \( N \).
### Table I: Length estimates

<table>
<thead>
<tr>
<th></th>
<th>( n_1 )</th>
<th>( n_2 )</th>
<th>( n )</th>
<th>( N )</th>
<th>( \hat{N}_H )</th>
<th>( \hat{N} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>CACM 1</td>
<td>10</td>
<td>18</td>
<td>28</td>
<td>104</td>
<td>108</td>
<td>107</td>
</tr>
<tr>
<td>CACM 2</td>
<td>14</td>
<td>8</td>
<td>22</td>
<td>84</td>
<td>77</td>
<td>76</td>
</tr>
<tr>
<td>CACM 3</td>
<td>18</td>
<td>41</td>
<td>59</td>
<td>454</td>
<td>294</td>
<td>288</td>
</tr>
<tr>
<td>CACM 4</td>
<td>16</td>
<td>21</td>
<td>37</td>
<td>137</td>
<td>156</td>
<td>155</td>
</tr>
<tr>
<td>CACM 5</td>
<td>15</td>
<td>16</td>
<td>31</td>
<td>124</td>
<td>123</td>
<td>122</td>
</tr>
<tr>
<td>CACM 6</td>
<td>15</td>
<td>13</td>
<td>28</td>
<td>99</td>
<td>107</td>
<td>107</td>
</tr>
<tr>
<td>CACM 7</td>
<td>10</td>
<td>9</td>
<td>19</td>
<td>59</td>
<td>62</td>
<td>61</td>
</tr>
<tr>
<td>CACM 8</td>
<td>15</td>
<td>16</td>
<td>31</td>
<td>133</td>
<td>123</td>
<td>122</td>
</tr>
<tr>
<td>CACM 9</td>
<td>19</td>
<td>41</td>
<td>60</td>
<td>312</td>
<td>300</td>
<td>294</td>
</tr>
<tr>
<td>CACM 10</td>
<td>9</td>
<td>9</td>
<td>18</td>
<td>48</td>
<td>57</td>
<td>57</td>
</tr>
<tr>
<td>CACM 11</td>
<td>9</td>
<td>9</td>
<td>18</td>
<td>55</td>
<td>57</td>
<td>57</td>
</tr>
<tr>
<td>CACM 12</td>
<td>11</td>
<td>9</td>
<td>20</td>
<td>62</td>
<td>67</td>
<td>66</td>
</tr>
<tr>
<td>CACM 13</td>
<td>11</td>
<td>9</td>
<td>19</td>
<td>61</td>
<td>62</td>
<td>61</td>
</tr>
<tr>
<td>CACM 14</td>
<td>15</td>
<td>25</td>
<td>40</td>
<td>187</td>
<td>174</td>
<td>173</td>
</tr>
<tr>
<td>GM 15</td>
<td>27</td>
<td>100</td>
<td>127</td>
<td>686</td>
<td>793</td>
<td>760</td>
</tr>
<tr>
<td>GM 28</td>
<td>49</td>
<td>214</td>
<td>263</td>
<td>1919</td>
<td>1931</td>
<td>1851</td>
</tr>
<tr>
<td>GM 36</td>
<td>47</td>
<td>329</td>
<td>376</td>
<td>2642</td>
<td>3012</td>
<td>2841</td>
</tr>
<tr>
<td>GM 40</td>
<td>82</td>
<td>433</td>
<td>515</td>
<td>3985</td>
<td>4317</td>
<td>4124</td>
</tr>
<tr>
<td>GM 50</td>
<td>35</td>
<td>168</td>
<td>203</td>
<td>1248</td>
<td>1421</td>
<td>1353</td>
</tr>
<tr>
<td>GM 118</td>
<td>13</td>
<td>24</td>
<td>37</td>
<td>122</td>
<td>158</td>
<td>155</td>
</tr>
</tbody>
</table>
The mean $\Delta$ of the relative errors $(N - \hat{N})/N$ yields

$$\Delta = \frac{1}{20} \sum (N - \hat{N})/N = -0.020,$$

which is smaller than the mean $\Delta_H$ of the relative errors $(N - \hat{N}_H)/N$,

$$\Delta_H = \frac{1}{20} \sum (N - \hat{N}_H)/N = -0.041.$$

Furthermore, the mean $\delta$ of the squared relative errors of $\hat{N}$ and $\delta_H$ of $\hat{N}_H$ shows a similar pattern:

$$\delta = \frac{1}{20} \sum ((N - \hat{N})/N)^2 = 0.016,$$
$$\delta_H = \frac{1}{20} \sum ((N - \hat{N}_H)/N)^2 = 0.019.$$

These examples show convincingly that the estimate $\hat{N}$ is slightly better than $\hat{N}_H$. The property of $\hat{N}$ of being more simple than $\hat{N}_H$ is an additional argument to prefer $\hat{N}$. 
III. Relationship between \( n^* \) and \( n \)

The equations

\[
V = (V^*)^2 / \lambda = (n^* \ln n^*)^2 / \lambda \quad (4)
\]

and

\[
V = N \ln n \approx \hat{N} \ln n
= n \ln (n/2) \ln n
= n \ln(n/\sqrt{2})^2
\quad (5)
\]

can be used to establish a simple relationship between \( n^* \) and \( n \). From (4) and (5), we get

\[
n^* \ln n^* = \sqrt{\lambda n} \ln (n/\sqrt{2}).
\]

This can be written as

\[
H = n^* \ln n^* - a \ln w = 0
\quad (6)
\]

with \( w = \sqrt{n/\sqrt{2}} \) and \( a = \sqrt{4 \lambda \sqrt{2}} \). We solve (8) approximately by applying the Newton-Raphson method to the first approximation \( w_0 = n^*/a \):

\[
w_1 = w_0 - (H(w_0) / H'(w_0))
= (n^*/a) + (n^* \ln n^* - n^* \ln (n^*/a)) / (a \ln (e n^*/a)).
\]

If we replace \( w \) and \( a \), we have \( n \) as a function of \( n^* \) or, more accurately, we have an approximative value \( \hat{n} \) of \( n \):

\[
\hat{n} = \frac{(n^*)^2}{4 \lambda} \left( \frac{1 + \ln n^*}{1 - \ln \sqrt{4 \lambda \sqrt{2} + \ln n^*}} \right)^2
\quad (7)
= \frac{(n^*)^2}{4 \lambda} \cdot h^n(n^*, \lambda).
\]

It may be noted that \( h(n^*, \lambda) \to 1 \) for \( n^* \to \infty \).

The following table shows some values of \( \hat{n} \):

<table>
<thead>
<tr>
<th>( n^* )</th>
<th>( \lambda )</th>
<th>1</th>
<th>1.25</th>
<th>1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>14</td>
<td>13</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>46</td>
<td>40</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>921</td>
<td>780</td>
<td>681</td>
<td></td>
</tr>
</tbody>
</table>
The values of this table correspond very well to the values of figure 10.3 in [1]. Equation (7) may be used to express \( N = n \log(n/2) \) as a function of \( n^* \) and \( \lambda \).

Of course, using (6) and applying the same methods we can express \( n^* \) as a function of \( n \) and \( \lambda \) and obtain the approximation

\[
\hat{n}^* = 2\sqrt{\lambda n} \cdot \left( \frac{1 + \ln \frac{n}{\sqrt{2}}}{1 + \ln \frac{4\sqrt{n}/\sqrt{2}}{\sqrt{2} + \ln \sqrt{n}/\sqrt{2}}} \right),
\]

where we have \( \hat{n}^* \rightarrow 2\sqrt{\lambda n} \) for \( n \rightarrow \infty \).
IV. Boundary length estimate

Formula (2) as well as formula (3) have the deficiency to be inaccurate for the smallest possible value of n, i.e. \(n = n^*\). In this case the length \(N = n^*\) whereas we get \(\hat{N} = n^* \text{ld}(n^*/2)\). In order to correct this deficiency we replace \(\hat{N} = n \text{ld}(n/2)\) by

\[
\hat{N}_B = n \text{ld}\left(\sqrt{n^2 - \left(\frac{n^*}{2}\right)^2} + 4\right) = n(\text{ld} n - 1 + \frac{1}{2} \text{ld}(1 + (16-n^*^2)/n^2)) = n(\text{ld} n - 1 + r(n,n^*)).
\]

where

\[
r(n,n^*) = \frac{1}{2} \text{ld}(1 + (16-n^*^2)/n^2)).
\]

This replacement can be interpreted as follows: The linear function \(n/2\) is approximated by the hyperbola \(\sqrt{n^2 - (n^*/2)^2} + 4\) whose asymptotic behavior corresponds to the linear function \(n/2\). But, for \(n = n^*\) we get \(\hat{N}_B = n^*\) now as required.

Table II shows that \(\hat{N}_B\) is a very good approximation of \(N\). The mean \(\Delta_B\) of the relative errors \((N - \hat{N}_B)/N\) yields

\[
\Delta_B = \frac{1}{20} \sum (N - \hat{N}_B)/N = -0.015,
\]

and the mean \(\delta_B\) of the squared relative errors gives

\[
\delta_B = \frac{1}{20} \sum ((N - \hat{N}_B)/N)^2 = 0.016.
\]

Thus, \(\hat{N}_B\) is as good as \(\hat{N}\) or \(\hat{N}_H\).

We remark:

a) In (9), it is \(r(n,n^*) \to 0\) for \(n \to \infty\) since, by (8), \(n^* \to \sqrt{n}\).

b) \(\hat{N}_B \leq \hat{N} \leq \hat{N}_H\) for \(n^* \geq 4\).

c) The notion of the boundary program volume [1] becomes obsolete if we use (9) as estimate for \(N\).
Table II: Boundary length estimate

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>$n^*$</th>
<th>N</th>
<th>$\hat{N}_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CACM 1</td>
<td>28</td>
<td>10</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>CACM 2</td>
<td>22</td>
<td>5</td>
<td>84</td>
<td>76</td>
</tr>
<tr>
<td>CACM 3</td>
<td>59</td>
<td>14</td>
<td>454</td>
<td>286</td>
</tr>
<tr>
<td>CACM 4</td>
<td>37</td>
<td>7</td>
<td>137</td>
<td>155</td>
</tr>
<tr>
<td>CACM 5</td>
<td>31</td>
<td>6</td>
<td>124</td>
<td>122</td>
</tr>
<tr>
<td>CACM 6</td>
<td>28</td>
<td>6</td>
<td>99</td>
<td>106</td>
</tr>
<tr>
<td>CACM 7</td>
<td>19</td>
<td>5</td>
<td>59</td>
<td>61</td>
</tr>
<tr>
<td>CACM 8</td>
<td>31</td>
<td>7</td>
<td>133</td>
<td>122</td>
</tr>
<tr>
<td>CACM 9</td>
<td>60</td>
<td>12</td>
<td>312</td>
<td>293</td>
</tr>
<tr>
<td>CACM 10</td>
<td>18</td>
<td>5</td>
<td>48</td>
<td>57</td>
</tr>
<tr>
<td>CACM 11</td>
<td>18</td>
<td>5</td>
<td>55</td>
<td>57</td>
</tr>
<tr>
<td>CACM 12</td>
<td>20</td>
<td>5</td>
<td>62</td>
<td>66</td>
</tr>
<tr>
<td>CACM 13</td>
<td>19</td>
<td>5</td>
<td>61</td>
<td>61</td>
</tr>
<tr>
<td>CACM 14</td>
<td>40</td>
<td>9</td>
<td>187</td>
<td>172</td>
</tr>
<tr>
<td>GM 15</td>
<td>127</td>
<td>34</td>
<td>686</td>
<td>754</td>
</tr>
<tr>
<td>GM 28</td>
<td>263</td>
<td>45</td>
<td>1919</td>
<td>1846</td>
</tr>
<tr>
<td>GM 36</td>
<td>375</td>
<td>70</td>
<td>2642</td>
<td>2831</td>
</tr>
<tr>
<td>GM 40</td>
<td>515</td>
<td>60</td>
<td>3985</td>
<td>4119</td>
</tr>
<tr>
<td>GM 60</td>
<td>203</td>
<td>34</td>
<td>1248</td>
<td>1349</td>
</tr>
<tr>
<td>GM 118</td>
<td>37</td>
<td>8</td>
<td>122</td>
<td>155</td>
</tr>
</tbody>
</table>
References
