
Subject: Re: Efficient comparison of arrays

Posted by [David Foster](#) on Wed, 13 Aug 1997 07:00:00 GMT

[View Forum Message](#) <> [Reply to Message](#)

J.D. Smith wrote:

```
>
> Just to keep the Astronomy department here from being one-upped....
>
> <SNIP>
>
> Here is an implementation I just made up:
>
> function contain,a,b
>   flag=[replicate(0b,n_elements(a)),replicate(1b,n_elements(b) )]
>   s=[a,b]
>   srt=sort(s)
>   s=s(srt) & flag=flag(srt)
>   wh=where(s eq shift(s,-1) and flag ne shift(flag, -1),cnt)
>   if cnt ne 0 then return, s[wh]
>   return,-1
> end
>
> I ran some time tests on the two implementations. While a_in_b is
> adequate for small vectors, it is prohibitively slow for large ones. An
> example averages the result in seconds for two 10000 element random
> integer vectors on the range [0,20000].
>
> Results for a_in_b:
>   Average Time:      19.669667
>
> Results for contain:
>   Average Time:      0.19233332
>
> Ratio:                102.269
>
> And for 100 element vectors in the range [0,200]:
>
> Results for a_in_b:
>   Average Time:      0.010666664
>
> Results for contain:
>   Average Time:      0.0015666644
>
```

When you choose a method make sure you test the solutions on data that is typical to your operations; don't rely on time postings based on artificial situations. Below are results comparing FIND_ELEMENTS.PRO (my routine that I've posted already)

and JD Smith's CONTAIN.PRO function listed above.

The test data are:

A = BYTARR(65536)

A 256x256 image which is a section of the brain that has been coded into discrete values to represent the different structures in the brain. Roughly in the range 0-128, many repeated values (compresses well). Very typical for my situation.

B = BINDGEN(50)

Here are the results:

```
IDL> t1=systime(1) & c = FIND_ELEMENTS(a,b) & t2=systime(1) & $
print, t2-t1
    2.3154050
IDL> t1=systime(1) & d = CONTAIN(a,b) & t2=systime(1) & $
print, t2-t1
    132.54824
```

In some situations the more primitive approach may be better (JD Smith's solution is certainly much more elegant and clever). Also be aware that some solutions like FIND_ELEMENTS() and WHERE_ARRAY() return *all* subscripts for items found, including repeats, whereas CONTAIN() does not.

Dave

--

~~~~~  
David S. Foster      Univ. of California, San Diego  
Programmer/Analyst   Brain Image Analysis Laboratory  
foster@bial1.ucsd.edu   Department of Psychiatry  
(619) 622-5892      8950 Via La Jolla Drive, Suite 2200  
                         La Jolla, CA 92037  
~~~~~