Subject: Re: What? You can't histogram a string array?
Posted by JD Smith on Mon, 27 Nov 2006 21:56:23 GMT
View Forum Message <> Reply to Message

On Mon, 27 Nov 2006 10:26:20 -0800, Braedley wrote:

> I'm very disappointed. I had a beautiful solution to a problem which
> involved determining if all the elements in one array exist in a second
> using histogram, but apparently I can't do that with string arrays. Oh
> well, I think I've seen something else in the built in library that'll
> do it just as fast and easily.

http://www.dfanning.com/tips/set_operations.html

ind_int_SORT is probably what you want.

(Has anyone else noticed that most reply posts these days start by
referencing some dfanning.com link?)

JD

Subject: Re: What? You can't histogram a string array?
Posted by David Fanning on Mon, 27 Nov 2006 22:11:09 GMT
View Forum Message <> Reply to Message

JD Smith writes:

> (Has anyone else noticed that most reply posts these days start by
> referencing some dfanning.com link?)

Maybe I need to implement better search technology. :-(

This site has expanded quite a ways beyond my original
vision for it. If anyone has suggestions for how it might
be better organized, and the suggestion doesn't take man-years
to implement, I'm all ears.

Cheers,

David
--
David Fanning, Ph.D.
Fanning Software Consulting, Inc.
Coyote's Guide to IDL Programming: http://www.dfanning.com/
Sepore ma de ni thui. ("Perhaps thou speakest truth.")

Subject: Re: What? You can't histogram a string array?
Posted by news.qwest.net on Mon, 27 Nov 2006 23:28:49 GMT
View Forum Message <> Reply to Message

"David Fanning" <news@dfanning.com> wrote in message
news:MPG.1fd510686e7366e2989de7@news.frii.com...
> JD Smith writes:
>
>> (Has anyone else noticed that most reply posts these days start by
>> referencing some dfanning.com link?)
>
> Maybe I need to implement better search technology. :-(
>
> This site has expanded quite a ways beyond my original
> vision for it. If anyone has suggestions for how it might
> be better organized, and the suggestion doesn't take man-years
> to implement, I'm all ears.

I think you have the optimal search alrgorithm as it stands now.
1)user posts to comp.lang.idl-pvwave
2) read response with link to dfanning.com page
3) click on link

It is very effective.

Subject: Re: What? You can't histogram a string array?
Posted by Braedley on Tue, 28 Nov 2006 12:41:24 GMT
View Forum Message <> Reply to Message

David Fanning wrote:
> JD Smith writes:
>
>> (Has anyone else noticed that most reply posts these days start by
>> referencing some dfanning.com link?)
>
> Maybe I need to implement better search technology. :-(
>
> This site has expanded quite a ways beyond my original
> vision for it. If anyone has suggestions for how it might
> be better organized, and the suggestion doesn't take man-years
> to implement, I'm all ears.
>
> Cheers,
>
> David
> --
> David Fanning, Ph.D.

> Fanning Software Consulting, Inc.
> Coyote's Guide to IDL Programming: http://www.dfanning.com/
> Sepore ma de ni thui. ("Perhaps thou speakest truth.")

The ironic thing is that I have read this article in the past, but
forgot which section it was in.

Braedley

Subject: Re: What? You can't histogram a string array?
Posted by JD Smith on Tue, 28 Nov 2006 16:31:47 GMT
View Forum Message <> Reply to Message

On Mon, 27 Nov 2006 16:28:49 -0700, R.G. Stockwell wrote:

>
> "David Fanning" <news@dfanning.com> wrote in message
> news:MPG.1fd510686e7366e2989de7@news.frii.com...
>> JD Smith writes:
>>
>>> (Has anyone else noticed that most reply posts these days start by
>>> referencing some dfanning.com link?)
>>
>> Maybe I need to implement better search technology. :-(
>>
>> This site has expanded quite a ways beyond my original
>> vision for it. If anyone has suggestions for how it might
>> be better organized, and the suggestion doesn't take man-years
>> to implement, I'm all ears.
>
> I think you have the optimal search alrgorithm as it stands now.
> 1)user posts to comp.lang.idl-pvwave
> 2) read response with link to dfanning.com page
> 3) click on link
>
> It is very effective.

Yeah, my point was a positive one. Without your site, we'd be reduced to
"Search the archive for something we may have written regarding this topic
last year or maybe before that", instead of pointing to a nicely
formatted page with oft-humorous editorial introductory notes. An
excellent resource whose value grows day by day. Thanks for keeping it up,
David. I for one try to click on one of your ads of interest when I visit
to keep the hosting fees covered.

JD

Subject: Re: What? You can't histogram a string array?
Posted by David Fanning on Tue, 28 Nov 2006 16:47:31 GMT
View Forum Message <> Reply to Message

JD Smith writes:

> I for one try to click on one of your ads of interest when I visit
> to keep the hosting fees covered.

Well, if you and ten thousand of your friends keep this up,
it's possible I may be able to buy a case or two of beer
for the next IEPA gathering. :-)

Cheers,

David

P.S. I *really* appreciate it, though!

--
David Fanning, Ph.D.
Fanning Software Consulting, Inc.
Coyote's Guide to IDL Programming: http://www.dfanning.com/
Sepore ma de ni thui. ("Perhaps thou speakest truth.")

Subject: Re: What? You can't histogram a string array?
Posted by Braedley on Tue, 28 Nov 2006 17:16:12 GMT
View Forum Message <> Reply to Message

JD, a small nitpick: ind_int_sort will occasionally take the index from
[a, b], and not from just a. This can quickly lead to out of bounds
conditions if the user doesn't want to index [a, b], but just wants to
index a. In my case, a is a column from a 2D string array, where b is
just a 1D string array. I think a where statement is all that is
needed to fix this (I know, it'll slow it down for large sets).

Braedley

Subject: Re: What? You can't histogram a string array?
Posted by news.qwest.net on Tue, 28 Nov 2006 17:19:15 GMT
View Forum Message <> Reply to Message

"David Fanning" <news@dfanning.com> wrote in message
news:MPG.1fd6160cadb472da989def@news.frii.com...
> JD Smith writes:
>

>> I for one try to click on one of your ads of interest when I visit
>> to keep the hosting fees covered.
>
> Well, if you and ten thousand of your friends keep this up,
> it's possible I may be able to buy a case or two of beer
> for the next IEPA gathering. :-)

Wow, I must admit I have not noticed the ads before, how long
have they been around? I'll have to start paying attention to them.
(And perhaps assign clicking duties to some underlings)

Cheers,
bob

Subject: Re: What? You can't histogram a string array?
Posted by David Fanning on Tue, 28 Nov 2006 17:27:37 GMT
View Forum Message <> Reply to Message

R.G. Stockwell writes:

> Wow, I must admit I have not noticed the ads before, how long
> have they been around?

Since the unemployment insurance ran out. :-(

They bring in a solid dollar a day, rain or shine.
It's just about enough to keep the site on the air.

Cheers,

David

--
David Fanning, Ph.D.
Fanning Software Consulting, Inc.
Coyote's Guide to IDL Programming: http://www.dfanning.com/
Sepore ma de ni thui. ("Perhaps thou speakest truth.")

Subject: Re: What? You can't histogram a string array?
Posted by Braedley on Tue, 28 Nov 2006 17:52:06 GMT
View Forum Message <> Reply to Message

Braedley wrote:
> JD, a small nitpick: ind_int_sort will occasionally take the index from

> [a, b], and not from just a. This can quickly lead to out of bounds
> conditions if the user doesn't want to index [a, b], but just wants to
> index a. In my case, a is a column from a 2D string array, where b is
> just a 1D string array. I think a where statement is all that is
> needed to fix this (I know, it'll slow it down for large sets).
>
> Braedley

Actually, the fix was much easier than previously thought. Instead of
return, srt[wh]
use
return, srt[wh]<srt[wh+1]

I haven't done any tests, but it shouldn't take much longer for sparse
or small sets.

Braedley

Subject: Re: What? You can't histogram a string array?
Posted by JD Smith on Tue, 28 Nov 2006 18:08:24 GMT
View Forum Message <> Reply to Message

On Tue, 28 Nov 2006 09:16:12 -0800, Braedley wrote:

> JD, a small nitpick: ind_int_sort will occasionally take the index from
> [a, b], and not from just a. This can quickly lead to out of bounds
> conditions if the user doesn't want to index [a, b], but just wants to
> index a. In my case, a is a column from a 2D string array, where b is
> just a 1D string array. I think a where statement is all that is
> needed to fix this (I know, it'll slow it down for large sets).

This is not good, and much worse than a minor nitpick. The
IND_INT_SORT algorithm relies on SORT doing the right thing. That is,
for two identical elements in the concatenated vector [a,b], SORT
should place the first one first, i.e. the matching elements from 'a'
will show up before those from 'b'. That's the only reason it
works. There was always the concern that IDL's SORT would change and
this would no longer be the case (the element from b would come
first), in which case the algorithm would be broken.

Can you provide an example where this isn't happening? I just tried
it on a simulated set of 100,000 random 6 character strings, and it
didn't show this behavior: all ~30 matching elements were selected
from a. I then ran this test 100 times, and in all cases it behaved
as expected. Perhaps it depends on the machine/OS? I'm actually not
sure if SORT calls a library sort function (which might make the
algorithm non-portable), or uses its own. You can try this test

yourself, like this:

```
for i=1,100 do begin
  a=string(byte(randomu(sd,6,100000)*26)+65b)
  b=string(byte(randomu(sd,6,100000)*26)+65b)
  s=ind_int_sort(a,b)
  print,strtrim(n_elements(s),2),' matches found'
  m=max(s)
  if m ge 100000 then begin
    print,'Out of bounds: ',m
    break
  endif
endfor
```

Let me know if it runs through without error for you. For anyone else who wants to test this, it would be appreciated. Here I run:

```
IDL> help,!VERSION,/st
** Structure !VERSION, 8 tags, length=76, data length=76:
   ARCH            STRING    'x86'
   OS              STRING    'linux'
   OS_FAMILY       STRING    'unix'
   OS_NAME         STRING    'linux'
   RELEASE         STRING    '6.3'
   BUILD_DATE      STRING    'Mar 23 2006'
   MEMORY_BITS     INT             32
   FILE_OFFSET_BITS
                   INT             64
```

BTW, if you only want the *values*, not the positions, where match occurred, replace:

```
return,srt[wh]
```

with

```
return,s[wh]
```

and this will "solve" the problem for you (with this change, it's equivalent to the CONTAIN function I posted long long ago). This is insensitive to the ordering of a or b SORT performs.

Also note that IND_INT_SORT only returns *one* match for repeated elements, which may or may not be what you want.

JD

Subject: Re: What? You can't histogram a string array?
Posted by JD Smith on Tue, 28 Nov 2006 18:12:53 GMT
View Forum Message <> Reply to Message

On Tue, 28 Nov 2006 09:52:06 -0800, Braedley wrote:

>
> Braedley wrote:
>> JD, a small nitpick: ind_int_sort will occasionally take the index from
>> [a, b], and not from just a. This can quickly lead to out of bounds
>> conditions if the user doesn't want to index [a, b], but just wants to
>> index a. In my case, a is a column from a 2D string array, where b is
>> just a 1D string array. I think a where statement is all that is
>> needed to fix this (I know, it'll slow it down for large sets).
>>
>> Braedley
>
> Actually, the fix was much easier than previously thought. Instead of
> return, srt[wh]
> use
> return, srt[wh]<srt[wh+1]
>
> I haven't done any tests, but it shouldn't take much longer for sparse
> or small sets.

That is a clever fix, but if the ordering of elements from a and b is
random, and if you have a repeated set in a match a repeated set in b, and
their interleaved sorted order is random, you'll get back a random number
of the matching repeats (not 1, as was intended).

See my other post though, and let me know your findings w.r.t. SORT.

Thanks,

JD

Subject: Re: What? You can't histogram a string array?
Posted by David Fanning on Tue, 28 Nov 2006 18:43:20 GMT
View Forum Message <> Reply to Message

JD Smith writes:

> This is not good, and much worse than a minor nitpick. The
> IND_INT_SORT algorithm relies on SORT doing the right thing. That is,
> for two identical elements in the concatenated vector [a,b], SORT
> should place the first one first, i.e. the matching elements from 'a'
> will show up before those from 'b'. That's the only reason it

> works. There was always the concern that IDL's SORT would change and
> this would no longer be the case (the element from b would come
> first), in which case the algorithm would be broken.

In running the test program, I get immediate out-of-bounds
errors with IDL's SORT routine. But nothing of the sort
(a pun!) with the NASA BSORT routine I always use when I
need to sort something "for real".

Running on Windows XP.

Cheers,

David
--
David Fanning, Ph.D.
Fanning Software Consulting, Inc.
Coyote's Guide to IDL Programming: http://www.dfanning.com/
Sepore ma de ni thui. ("Perhaps thou speakest truth.")

Subject: Re: What? You can't histogram a string array?
Posted by David Fanning on Tue, 28 Nov 2006 18:49:50 GMT
View Forum Message <> Reply to Message

David Fanning writes:

> In running the test program, I get immediate out-of-bounds
> errors with IDL's SORT routine. But nothing of the sort
> (a pun!) with the NASA BSORT routine I always use when I
> need to sort something "for real".

Whoops! I forgot the mandatory link:

http://www.dfanning.com/tips/sort.html

Cheers,

David
--
David Fanning, Ph.D.
Fanning Software Consulting, Inc.
Coyote's Guide to IDL Programming: http://www.dfanning.com/
Sepore ma de ni thui. ("Perhaps thou speakest truth.")

Subject: Re: What? You can't histogram a string array?

Posted by news.qwest.net on Tue, 28 Nov 2006 18:53:43 GMT
View Forum Message <> Reply to Message

"David Fanning" <news@dfanning.com> wrote in message
news:MPG.1fd632bc5a98f848989df5@news.frii.com...
> David Fanning writes:
>
>> In running the test program, I get immediate out-of-bounds
>> errors with IDL's SORT routine. But nothing of the sort
>> (a pun!) with the NASA BSORT routine I always use when I
>> need to sort something "for real".
>
> Whoops! I forgot the mandatory link:
>
> http://www.dfanning.com/tips/sort.html
>

You know, those ad links are actually pretty good. I'm looking
at the Princeton Instruments brochure right now, to see the latest
in near infrared imaging systems. I wonder if the tech has gotten to
the point where we can use off the shelf stuff now, instead of building
our own.

Subject: Re: What? You can't histogram a string array?
Posted by Braedley on Tue, 28 Nov 2006 19:23:58 GMT
View Forum Message <> Reply to Message

JD Smith wrote:
> On Tue, 28 Nov 2006 09:52:06 -0800, Braedley wrote:
>
>>
>> Braedley wrote:
>>> JD, a small nitpick: ind_int_sort will occasionally take the index from
>>> [a, b], and not from just a. This can quickly lead to out of bounds
>>> conditions if the user doesn't want to index [a, b], but just wants to
>>> index a. In my case, a is a column from a 2D string array, where b is
>>> just a 1D string array. I think a where statement is all that is
>>> needed to fix this (I know, it'll slow it down for large sets).
>>>
>>> Braedley
>>
>> Actually, the fix was much easier than previously thought. Instead of
>> return, srt[wh]
>> use
>> return, srt[wh]<srt[wh+1]
>>
>> I haven't done any tests, but it shouldn't take much longer for sparse

>> or small sets.
>
> That is a clever fix, but if the ordering of elements from a and b is
> random, and if you have a repeated set in a match a repeated set in b, and
> their interleaved sorted order is random, you'll get back a random number
> of the matching repeats (not 1, as was intended).
>
> See my other post though, and let me know your findings w.r.t. SORT.
>
> Thanks,
>
> JD

I hit an out of bounds on my first try. Running MacOSX, 10.4.8,
IDLv6.2. Unfortunately, I do need the indices, as I pointed out
earlier. Perhaps I'll use BSORT instead.

Subject: Re: What? You can't histogram a string array?
Posted by JD Smith on Tue, 28 Nov 2006 21:17:33 GMT
View Forum Message <> Reply to Message

On Tue, 28 Nov 2006 11:43:20 -0700, David Fanning wrote:

> JD Smith writes:
>
>> [quoted text muted]
>
> In running the test program, I get immediate out-of-bounds
> errors with IDL's SORT routine. But nothing of the sort
> (a pun!) with the NASA BSORT routine I always use when I
> need to sort something "for real".

OK, so far OSX and Windows XP throw out of bounds errors. Can anyone
on Linux confirm that this runs without error? I checked libidl.so,
and it mentions qsort, of the GLIBC variety. So it must be my
implementation of qsort in my GLIBC preserves order, but others do
not. Ouch.

Might want to add a note to that page. If you don't have repeated
elements, then the fix Braedley offered works fine. BSORT from
Nasalib sorts and then reorders duplicates to preserve the original
order. It will compromise speed somewhat, but is a good alternative.

JD

P.S. How long as it been the case that SORT scrambles order on Windows?
I'm surprised the issue with IND_INT_SORT didn't come up before.
