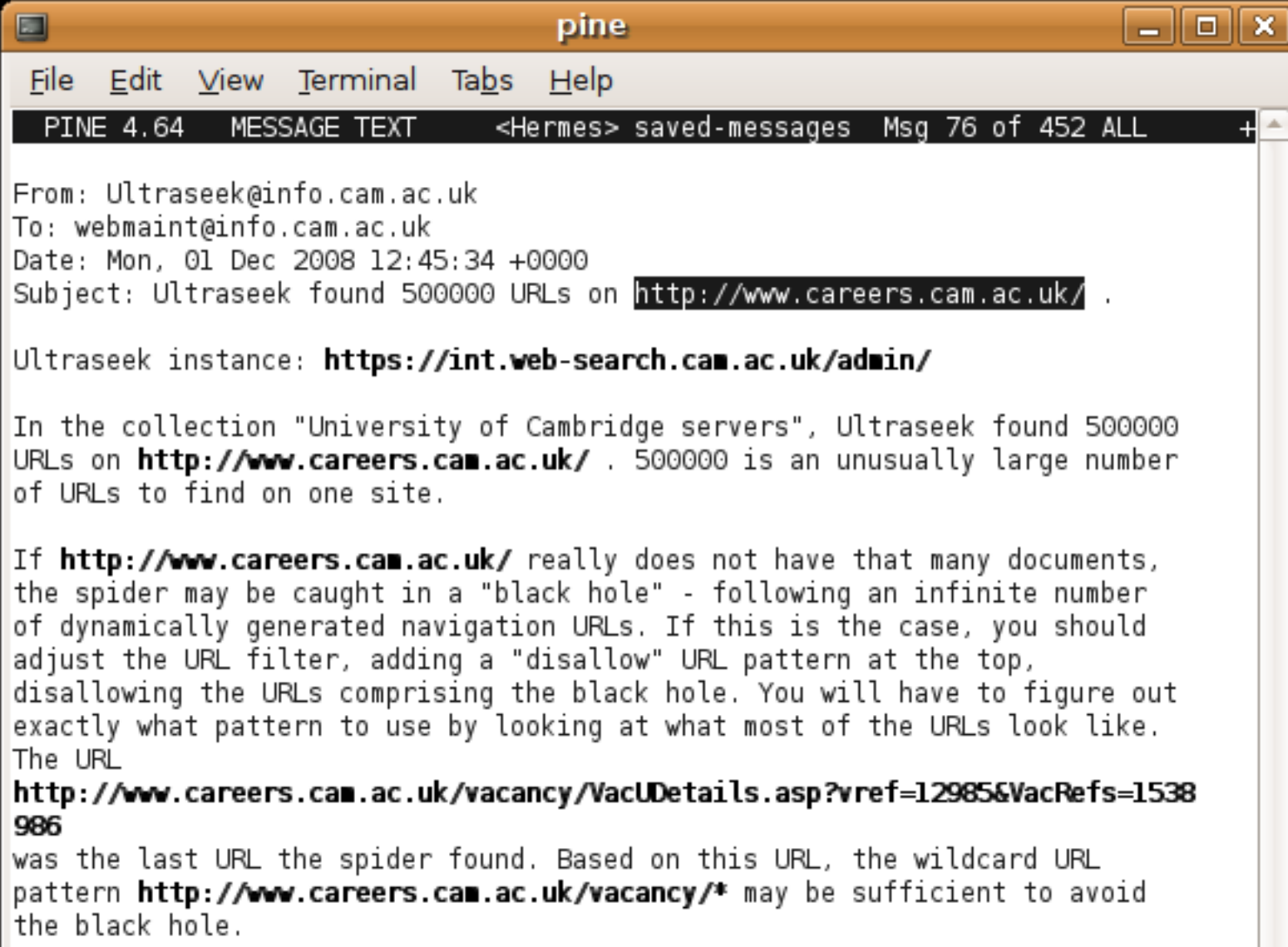


The first we know...



The screenshot shows a window titled "pine" with a menu bar (File, Edit, View, Terminal, Tabs, Help) and a status bar (PINE 4.64 MESSAGE TEXT <Hermes> saved-messages Msg 76 of 452 ALL). The email content is as follows:

```
From: Ultraseek@info.cam.ac.uk
To: webmaint@info.cam.ac.uk
Date: Mon, 01 Dec 2008 12:45:34 +0000
Subject: Ultraseek found 500000 URLs on http://www.careers.cam.ac.uk/ .

Ultraseek instance: https://int.web-search.cam.ac.uk/admin/

In the collection "University of Cambridge servers", Ultraseek found 500000
URLs on http://www.careers.cam.ac.uk/ . 500000 is an unusually large number
of URLs to find on one site.

If http://www.careers.cam.ac.uk/ really does not have that many documents,
the spider may be caught in a "black hole" - following an infinite number
of dynamically generated navigation URLs. If this is the case, you should
adjust the URL filter, adding a "disallow" URL pattern at the top,
disallowing the URLs comprising the black hole. You will have to figure out
exactly what pattern to use by looking at what most of the URLs look like.
The URL
http://www.careers.cam.ac.uk/vacancy/VacUDetails.asp?vref=12985&VacRefs=1538986
was the last URL the spider found. Based on this URL, the wildcard URL
pattern http://www.careers.cam.ac.uk/vacancy/\* may be sufficient to avoid
the black hole.
```

The problems

- Spiders are (must be) stupid
- Much off-the-shelf software doesn't expect spiders
- URLs vs. Documents
- URLs are case sensitive
- Same document, multiple URLs
- Identical documents that are not identical
 - Calendars, PATH_INFO
- Cache busting
- 'Status 200' error pages

The solutions

- “Well, don't do that then”
- URL validation
- robots.txt
- `<meta name="robots" content="...">`
 - “index” or “noindex”
 - “follow” or “nofollow”
 - “all”
 - “none”
- Filtering at the spider end