Redmine - Defect #2491

robots.txt file is has incorrect urls

2009-01-12 02:25 - Brad Schick

Status:	Closed	**Start date:**	2009-01-12
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:	0.9.0		
Resolution:	Fixed	**Affected version:**	

Description

My mongrel servers all crashed, I believe due to memory pressure from indexing robots. While looking into this I found that in Redmine 0.8 the public/robots.txt file has the following lines:

User-agent: *
Disallow: /projects/gantt
Disallow: /projects/calendar
Disallow: /repositories/diff

The gantt and calendar lines are invalid, however, and will not block robots. If you navigate to the calendar or gantt for a particular project the actual URLs are:

http://(host-name)/projects/(project-name)/issues/calendar
http://(host-name)/projects/(project-name)/issues/gantt

Since robots.txt does not support wild cards, perhaps the URLs should be modified to allow a stock robots.txt to work. Something more like repository URLs:

http://(host-name)/repositories/show/(project-name)

Associated revisions

Revision 2319 - 2009-01-27 18:27 - Jean-Philippe Lang

Replaces the obsolete robots.txt with a cached action (#2491).

Revision 2334 - 2009-01-29 14:54 - Jean-Philippe Lang

Adds rel='nofollow' attribute to other formats download links (#2491).

History

#1 - 2009-01-12 19:20 - Jean-Philippe Lang

Indeed, this file wasn't fixed when the URLs changed (eg. /projects/gantt/foo => /projects/foo/issues/gantt). And as you said, the robots.txt can not be fixed since wildcards are not supported.

In the future, all project related URLs should start with /projects/foo/. And I don't really want to change this just for this purpose.

What could be done:

- using wildcards: they are supported by Googlebot and Yahoo Slurp, the top 2 spiders here (maybe not the best solution)
- or adding an action that responds to /robots.txt and returns a valid file (no wildcards) for all visible projects.

What do you think?

#2 - 2009-01-12 20:31 - Brad Schick

It only takes one spider that doesn't accept wildcards to crash a site, so I don't think that is a good option.

Generating robots.txt sounds like a good idea (as long as it is cached). And at the risk of adding extra work, there could be settings to add/remove Redmine areas from it. That way people could simply check options like "repository", "repository diffs", "gantt charts", "wiki", etc. without having to think about their URLs.

#3 - 2009-01-12 21:11 - Jean-Philippe Lang

Sounds great.

Btw, I use mongrel for development only.
You should consider using another deployment solution, like mod_rails.
I personnaly run redmine.org without problems using mod_fcgid (bots traffic represents about 15GB/month here).

#4 - 2009-01-22 08:47 - Brad Schick

Exploring my site performance again I suspect that a fair amount of wasted CPU time is going into producing PDFs, CSV, and Atom representations on all issues, issue lists, wiki, forum pages, etc.

Along with the discussed changes for robots.txt, it would be very helpful to have options to add rel="nofollow" attributes to potentially expensive links like these. Here is a discussion of that attribute: http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=96569

#5 - 2009-01-27 18:42 - Jean-Philippe Lang

An action that responds to /robots.txt is added in r2319.

I'm not sure that rel="nofollow" attribute is appropriate here. Wikipedia says:

> How the attribute is being interpreted differs between the search engines. While some take it literally and do not follow the link to the page being linked to, others still "follow" the link to find new web pages for indexing. In the latter case rel="nofollow" actually tells a search engine "Don't score this link" rather than "Don't follow this link." This differs from the meaning of nofollow as used within a robots meta tag, which does tell a search engine: "Do not follow any of the hyperlinks in the body of this document."

Yahoo Slurp for example will actually follow the link.

#6 - 2009-01-27 19:38 - Brad Schick

Seems impractical to add every single one of those links to robots.txt (since it doesn't support wild cards). Are there other options? I think this is a real problem for my site.

#7 - 2009-01-29 15:00 - Jean-Philippe Lang

- Status changed from New to Closed

- Target version set to 0.9.0

- Affected version (unused) set to 0.8.0

- Resolution set to Fixed

rel=nofollow attribute is added in r2334. Spiders like Googlebot should no longer follow these links.
I close this ticket since the original request is fixed.

An other solution would be to look for common bots (using request's user-agent) to deny access to these links.
This could be done globally using a before_filter, maybe as a plugin.

#8 - 2009-01-29 17:05 - Jean-Philippe Lang

You can have a look at the BotsFilterPlugin.

#9 - 2009-08-09 17:51 - John Goerzen

rel=nofollow is not adequate. The bots may have already spidered all those links, and others can link in to them. A generated robots.txt would be better.