Indexer un blog, avec ses pages aux sujets variés

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 26 juin 2006

https://www.bortzmeyer.org/indexation-blog.html

Un truc très agaçant avec tous les moteurs de recherche est leur comportement lorsqu'une page contient des articles très divers et sans lien entre eux (deux cas typiques : les blogs et les archives de listes de diffusion). Le moteur voit tous les mots-clés sur la page et ne comprend pas les frontières entre articles. Ainsi, je vois (dans le journal du serveur Web) un malheureux arriver sur une page après avoir tapé "freebsd ldap" dans Google alors que la page a bien un article sur LDAP et un sur FreeBSD mais qu'ils ne sont pas reliés... Cela donne souvent des résultats surréalistes <http://mmdl.free.fr/blog-m/?p=330>.

Altavista avait l'opérateur NEAR qui aidait beaucoup à résoudre ce problème. Mais je ne connais aucun moteur actuel qui l'utilise.

Il semble que la solution (mal documentée <http://www.robotstxt.org/meta.html>) soit de mettre dans la section <head> de ses pages HTML, lorsqu'elles sont de type "navigation" (et contiennent donc des articles sans rapport entre eux) :

```
<meta name="robots" content="noindex,follow">
```

Cela semble bien fonctionner avec Google (qui le documente <http://googlewebmastercentral.blogspot.com/2007/03/using-robots-meta-tag.html>). Mais apparemment pas avec des concurrents comme Exalead.

J'utilise désormais ce <meta> sur toutes mes pages de navigation et je mets :

```
<meta name="robots" content="index,follow">
```

dans les pages ordinaires, celles qui contiennent un article et un seul.