Namazu, indexation de texte (mais pas au point)

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 6 novembre 2007

https://www.bortzmeyer.org/namazu.html

Cela fait longtemps que je me dis que je devrais indexer les giga-octets de textes ou de programmes qui occupent mon disque dur. Je viens de tester Namazu <http://www.namazu.org/>, mais qui ne convient pas.

Je cherche un programme simple, pas une usine à gaz nécessitant l'installation et la maintenance de plusieurs composants. Je ne suis pas convaincu de la nécessité d'installer et de maintenir un SGBD juste pour cette tâche. Autrefois, j'utilisais glimpse mais il était non-libre et semble de toute façon avoir complètement disparu, sans mainteneur.

Mes fichiers sont des documents LaTeX, Docbook, des sources de programme en divers langages, uniquement des formats ouverts, a priori relativement faciles à analyser (pas de MS-Word ni même d'obésiciels libres comme OpenOffice).

Je viens de tester Namazu <http://www.namazu.org/> mais c'est un échec. Namazu est un logiciel libre, développé par un programmeur japonais et dont la principale fonction mise en avant est la capacité à bien traiter les textes en japonais, une langue dont j'ignore tout.

Namazu est disponible facilement pour la plupart des systèmes. Sur Ubuntu, par exemple, c'est un paquetage existant. Une fois installé, on programme l'indexation, par exemple quotidienne. J'utilise cron et je crée donc un `/etc/cron.daily/index-blog` qui indexe, pour tester, une partie de mon disque dur (celle où je garde les fichiers de ce blog). Le fichier `index-blog` contient :

```
#!/bin/sh

HOME=/home/stephane

INDEX=$HOME/Index-Blog

if [ ! -d $INDEX ]; then
    mkdir $INDEX
fi

sudo -u stephane mknmz --all \
        --exclude '_darcs|\.(html|txt|tex|ps|pdf|rng|dvi|xml|atom|full_atom|bak)|~' \
        --meta --decode-base64 --check-filesize --output-dir=$INDEX $HOME/Blog
```

sudo est nécessaire car les tâches dans `cron.daily` sont exécutées par root. La liste des exclusions comporte les formats produits automatiquement, qu'il n'est pas utile d'indexer.

`mknmz` indexe donc le répertoire tous les jours et il le fait intelligemment. Seuls les nouveaux fichiers sont traités :

```
Looking for indexing files...
/home/stephane/Blog/RFC/2663-NOT-YET.rfc_xml was updated!
/home/stephane/Blog/RFC/3107-NOT-YET.rfc_xml was updated!
/home/stephane/Blog/RFC/3330.rfc_xml was updated!
/home/stephane/Blog/entries/signaler-a-signal-spam.entry_xml was updated!
/home/stephane/Blog/Makefile was updated!
14 files are found to be indexed.
...
```

Une fois les fichiers indexés, on peut chercher avec la commande `namazu`, à qui on doit indiquer le truc qu'on cherche et l'index :

```
% namazu TRUC ~/Index-Blog
 Total 12 documents matching your query.

1. namazu-NOT-YET.entry_xml (score: 2)
Author: unknown
Date: Sun, 21 Oct 2007 23:34:29 +0000
http://www.namazu.org/ Indexer avec exclusion ? Tester sur horcrux Voir ~/bin/index* namazu TRUC ~/Index-Blo
/home/stephane/Blog/entries/namazu-NOT-YET.entry_xml (876 bytes)
...
```

Namazu peut extraire des métadonnées de certains fichiers (comme la date ci-dessus) et permet les recherches via ces métadonnées.

Namazu est livré avec une grande quantité de **filtres** qui permettent d'analyser de nombreux formats (y compris les RFC). Un des ces filtres lit les courriers au format RFC 2822 [1] et, combiné avec les métadonnées, cela permet de chercher un message par son expéditeur, ici, les messages qui viennent de Bush :

```
% namazu +from:bush ~/Index-Mail
 Total 3 documents matching your query.
...
2. Re: anycast stability experiment (score: 1)
Author: Randy Bush <randy@psg.com>
Date: Thu, 24 Feb 2005 15:43:36 +0900
we promised to report results. well, there were fun events, such as northern-hemisphere winter holidays etc,
/homme/stephane/Mail/system/afnog-2005-02.gz (3,567 bytes)
```

Une autre commande pratique est `nmzgrep` qui lance grep sur les fichiers trouvés, donnant ainsi un affichage plus habituel :

1. Pour voir le RFC de numéro NNN, `https://www.ietf.org/rfc/rfcNNN.txt`, par exemple `https://www.ietf.org/rfc/rfc2822.txt`

```
% nmzgrep  'DB8' ~/Index-Blog
/home/stephane/Blog/entries/registre-temps-reel.entry_xml:
/home/stephane/Blog/entries/registre-temps-reel.entry_xml:          VALUES (1,'ns1.nic.example','2001:DB8::1035:
/home/stephane/Blog/RFC/3849.rfc_xml:<computer>2001:DB8::/32</computer> a été réservé et que les adresses
/home/stephane/Blog/RFC/5006.rfc_xml:<computer>2001:DB8:BEEF:42::/64</computer>). Le routeur diffuse ses
...
```

Mais Namazu ne semble pas encore très au point. Le nombre de bogues et le trafic quasi-nul sur la liste de diffusion semble indiquer que le logiciel n'a pas une communauté d'utilisateurs. Certains bogues sont vraiment très visibles, pourtant, comme le fait que, si on met plusieurs options identiques, les premières sont silencieusement ignorées, ou comme le cas de `--include`, qui prend en argument un nom de fichier et ne produit aucun message d'erreur si le fichier n'existe pas...

Chose plus étonnante, aucun filtre n'existe pour XML... Un de mes fichiers XSL a été pris pour du LaTeX simplement parce que le nom de ce système apparaissait dans un commentaire XML...

Les développeurs répondent bien sur la liste, mais dans un anglais pratiquement illisible. Eh oui, le monde est vaste et Babel est une réalité dès qu'on sort de quelques « élites » mondialisées.

Je continue donc mes recherches, sur d'autres logiciels. J'ai noté les candidats potentiels sur del.icio.us : `<http://del.icio.us/bortzmeyer/index>`.