RFC 6203 : IMAP4 Extension for Fuzzy Search

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 27 mars 2011

Date de publication du RFC : Mars 2011

https://www.bortzmeyer.org/6203.html

Où s'arrêtera la liste des extensions au protocole IMAP d'accès aux boîtes aux lettres ? En tout cas, elle ne cesse de grossir. Ce RFC propose une extension aux fonctions de recherche d'IMAP pour permettre des recherches floues, par exemple pour que « hipothèse » puisse trouver « hypothèse ».

Beaucoup de systèmes de recherche disposent de mécanisme analogue, par exemple le moteur de recherche Google, lorsqu'on lui soumet la requête erronée citée plus haut propose « Essayez avec cette orthographe : hypothèse 2 premiers résultats affichés ». Mais IMAP n'avait qu'un service de recherche exact, via la commande `SEARCH` (RFC 3501 [1], section 6.4.4). Notre RFC 6203 l'étend en permettant, via l'ajout du mot-clé `FUZZY`, des recherches floues. Il est important de noter que ce RFC ne spécifie qu'une syntaxe : la sémantique exacte d'une recherche est laissée à chaque mise en œuvre, et on peut penser qu'elle évoluera avec les progrès de la recherche floue. Ainsi, pour prendre un exemple, est-ce que « Penelope » trouvera les textes sur Pénélope ? Le RFC ne le spécifie pas, laissant chaque serveur IMAP ajouter les algorithmes de recherche floue qu'il préfère.

La syntaxe exacte figure en section 3. Un mot-clé, `FUZZY`, enregistré dans le registre IANA d'IMAP <https://www.iana.org/assignments/imap4-capabilities>, préfixe le critère de recherche, comme dans, par exemple :

```
C: MYTAG1 SEARCH FUZZY (SUBJECT "balade dames jadisse")
S: * SEARCH 1 5 10
S: MYTAG1 OK Search completed.
```

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc3501.txt

Et on trouvera peut-être (selon l'implémentation) un message dont le sujet était « La ballade des dames du temps jadis » (et merci à ma fille pour une intéressante discussion sur balade / ballade). Je répète que c'est le serveur qui décidera du niveau de flou acceptable, le RFC ne fournit pas de guide à ce sujet.

Le mot-clé `FUZZY` s'applique au critère qu'il préfixe, pas à toute la recherche. Donc :

```
C: MYTAG2 SEARCH FUZZY (SUBJECT "balade dames jadisse") FROM francois@villon.fr
S: * SEARCH 1 4
S: MYTAG2 OK Search completed.
```

fera une recherche floue sur le sujet mais une recherche exacte sur l'auteur.

Les recherches floues vont évidemment renvoyer davantage de résultats que les recherches exactes et le tri de ces résultats, par pertinence, est donc essentiel. La section 4 décrit donc le mécanisme de pertinence. Le RFC recommande très fortement aux serveurs IMAP qui mettent en œuvre `FUZZY` d'affecter une grandeur entre 0 et 100 à chaque résultat, indiquant la pertinence. Si le serveur accepte également l'extension `ESEARCH` (RFC 4731), celle-ci peut servir à récupérer les valeurs de pertinence, avec la nouvelle option de `RETURN`, `RELEVANCY` :

```
C: MYTAG3 SEARCH RETURN (RELEVANCY ALL) FUZZY TEXT "Bnjour"
S: * ESEARCH (TAG "MYTAG3") ALL 1,5,10 RELEVANCY (4 99 42)
S: MYTAG3 OK Search completed.
```

Les exemples ci-dessus utilisaient la recherche floue avec des chaînes de caractère. Mais la norme ne se limite pas à ces chaînes, comme le précise la section 5. Un serveur peut donc décider de permettre des recherches floues pour d'autres types de données et le RFC donne des indications sur les types pour lesquels c'est utile :

— Les dates, pour permettre des recherches comme « il y a une semaine, à peu près », la pertinence décroissant lorsqu'on s'éloigne de la date indiquée,
— Les tailles, pour chercher des messages d'« environ un Mo »,
— Alors qu'en revanche les recherches par UID (RFC 3501, section 2.3.1.1) n'ont sans doute aucune raison d'être floues.

Le concept de pertinence s'applique aussi à la commande `SORT` (section 6). On peut donc trier les messages selon la pertinence :

```
C: MYTAG4 SORT (RELEVANCY) UTF-8 FUZZY SUBJECT "Bjour"
S: * SORT 5 10 1
S: MYTAG4 OK Sort completed.
```

Si vous être programmeur et que vous ajoutez cette capacité de flou à un serveur IMAP existant, lisez bien la section 8 sur la sécurité. En effet, les recherches floues peuvent consommer bien plus de ressources du serveur et faciliter ainsi des attaques par déni de service.

D'autre part (comme, là encore, avec les moteurs de recherche du Web), les recherches floues peuvent être plus susceptibles d'empoisonnement par des méchants qui glisseraient dans les messages de quoi apparaître en tête du classement.

Et quelles sont les implémentations existantes de cette extension ? Il y en a apparemment au moins deux mais j'ai été trop paresseux pour chercher lesquelles, comptant sur les lecteurs pour me les signaler.