un algorithme pour supprimer les mots vides d'une requête
an algorithm to discard function words from a query

Jacques Vergne
GREYC Caen
avril 2008

Contexte :
dans le cadre de l'utilisation d'un moteur de recherche, la requête peut comporter plusieurs mots, et même une phrase, un paragraphe ou un texte

pourquoi supprimer les mots vides d'une requête ?
si on laisse les mots vides dans une requête, leur fréquence élevée perturbe le calcul de la fonction de ranking des documents de la réponse du moteur

Ressources : endogènes (les ressources sont dans les données à traiter)
aucune ressource externe : pas de dictionnaire ni de stop-list
pour chaque mot de la requête : sa longueur, et le nombre de documents indexés contenant ce mot (donnée permettant de comparer les fréquences de 2 mots, et accessible directement dans l'index du moteur)
donc l'algorithme est indépendant de la langue de la requête, et donc ne nécessite pas le diagnostic de la langue de la requête

Propriété exploitée : propriété **locale**
un mot vide est plus court (ou de même longueur) et plus fréquent que **ses voisins** (cf Zipf et Saussure)

(et cf mon article de CIDE 8 - mai 2005)

plus précisément, dans l'algorithme proposé, cette propriété est déclinée en 4 propriétés, exploitées dans un ordre quelconque :

1) un mot est vide s'il est plus court (ou de même longueur) et plus fréquent que ses voisins

2) le premier mot est vide s'il est plus court (ou de même longueur) et plus fréquent que le deuxième mot

3) 2 mots contigus forment un couple de 2 mots vides :
- si le premier est plus court (ou de même longueur) et plus fréquent que le mot précédent,
- et si le deuxième est plus court (ou de même longueur) et plus fréquent que le mot suivant,
- et si la différence des effectifs entre les 2 mots du couple (i et i+1) est inférieure à la différence avant le couple (i-1 et i) et inférieure à la différence après le couple (i+1 et i+2)

4) 3 mots contigus forment un triplet de 3 mots vides :
- si le premier est plus court (ou de même longueur) et plus fréquent que le mot précédent,
- et si le troisième est plus court (ou de même longueur) et plus fréquent que le mot suivant,
- et si la somme de la différence des effectifs entre les 2 premiers mots du triplet (i et i+1) et de la différence des effectifs entre les 2 derniers mots du triplet (i+1 et i+2) est inférieure à la différence avant le triplet (i-1 et i) et inférieure à la différence après le triplet (i+2 et i+3)

Chronologie de l'algorithme :
la requête est représentée sous la forme d'une liste de mots
les mots de la requête sont étiquetés vides ou non vides en exploitant les 4 propriétés susdites :
la propriété 1 en faisant glisser une fenêtre de 1 mot du début à la fin de la requête,
la propriété 3 en faisant glisser une fenêtre de 2 mots du début à la fin de la requête,
et la propriété 4 en faisant glisser une fenêtre de 3 mots du début à la fin de la requête ;
la propriété 2 ne concerne que les 2 premiers mots ;
les mots vides sont supprimés de la requête

Quelques exemples :

LeMonde_2008-01-15, 5000 documents indexés

```
 2  656 De
 9  574 nouvelles  *
10   91 techniques  *
10    4 facilitent  *
 2 4384 la
13   26 compréhension *
 2 4398 du
14   81 fonctionnement  *
 2 4398 du
 7   28 cerveau  *
```

nouvelles techniques facilitent compréhension fonctionnement cerveau

```
1 4241 l
2  204 or  *
2 4442 de
2 4384 la
6  106 Banque *
2 4442 de
6 1342 France  *
```

or Banque France

```
 5  770 Paris *
 2 4084 et
 6    6 Panama  *
 9   26 réclament  *
 1 4241 l
11   42 extradition  *
 2 4442 de
 7    7 Noriega  *
```

Paris Panama réclament extradition Noriega

Spiegel_2008-01-14, 1000 documents indexés

```
 3 737 Die
10   6 Nasa-Sonde  *
11   0 "Messenger"  *
 8  45 erreicht  *
 3 805 den
 6  20 Merkur  *
 1 993 -
 5 665 einen
 3 760 der
 9  15 heißesten  *
 4  26 Orte  *
 2 718 im
14   6 Planetensystem  *
```

Nasa-Sonde "Messenger" erreicht Merkur heißesten Orte Planetensystem

```
 4 223 Doch  *
 3 238 nun
 5  73 kehrt  *
 3 760 der
16   7 Ex-Außenminister  *
 3 566 wie
 5 111 zuvor  *
 5 328 schon
10  11 Altkanzler  *
 8  19 Schröder  *
 3 700 für
 5 665 einen
11   6 Kurzausflug  *
 3 457 ins
10  37 politische  *
 8  26 Geschäft  *
 6 210 zurück  *
```

Doch kehrt Ex-Außenminister zuvor Altkanzler Schröder Kurzausflug politische Geschäft zurück

HeraldTribune_2008-01-14, 1000 documents indexés

```
6 324 France  *
3 347 set  *
2 998 to
8 131 announce  *
3 131 UAE  *
7 179 nuclear  *
6 231 energy  *
4 233 deal  *
```

France set announce UAE nuclear energy deal

```
 7  59 Senator *
 4 604 John
 6 429 McCain *
 2 926 is
 3 443 now  *
 6  40 viewed *
 4 806 more
 9   8 favorably *
 4 718 than
 3 684 his
 5 417 major
11  25 competitors *
 5 367 while
 6 140 Barack *
 5 141 Obama  *
 3 923 has
 8  39 improved *
 3 684 his
 8  48 standing *
```

Senator McCain now viewed favorably competitors Barack Obama improved standing

Stampa_2008-01-14, 1000 documents indexés

```
2 695 La
9 148 posizione  *
3 800 del
4 148 Papa  *
2 393 su
7 124 Galileo  *
2 306 ci
6 117 umilia  *
1 762 e
2 306 ci  *
7 124 offende  *
```

posizione Papa Galileo umilia ci offende

```
 4 170 Dopo  *
 3 339 due
 7 125 denunce  *
 3 843 per
 5 124 abuso  *
 2 868 di
 9 148 posizione  *
 9 125 dominante  *
 9 540 Bruxelles  *
 5 150 mette  *
 5 249 sotto
 6 127 accusa  *
 1 504 l
16 119 interoperabilità  *
 1 762 e
 2 715 il
12 123 collegamento  *
 3 333 tra
 8 173 software  *
 8 127 separati  *
```

Dopo denunce abuso posizione dominante Bruxelles mette accusa interoperabilità collegamento software separati

ElPais_2008-01-14, 1000 documents indexés

```
 2 942 La
10  18 aceptación  *
 7  30 popular  *
 2 972 de
 2 958 la
 5  38 letra  *
 3 958 del
 5  30 himno  *
 2 549 ha
 4 194 sido  *
13  15 impresionante  *
```

aceptación popular letra himno sido impresionante

```
 3 356 Una
 6   7 viñeta *
 2 972 de
 6  28 Forges *
 7  79 refleja *
 3 188 muy
 4 137 bien *
 2 958 la
 9  99 situación *
 2 972 de
 2 958 la
15   7 intelectualidad *
 2 959 en
 6 954 España *
```

viñeta Forges refleja bien situación intelectualidad España

Kathimerini_2008-04-21, 1000 documents indexés

```
 7 131 Σύλληψη *
 3 474 δύο
 6 135 νεαρών *
 3 804 για
11 148 καλλιέργεια *
 7 131 ινδικής *
 8 130 κάνναβης *
 4 658 στις
 6 131 Σέρρες *
```

Σύλληψη νεαρών καλλιέργεια ινδικής κάνναβης Σέρρες

```
11  89 Προσπάθειες *
 3 804 για
12 117 απελευθέρωση *
 2 127 25
 8  88 ναυτικών *
 9  89 ισπανικού *
10  88 αλιευτικού *
 3 702 στη
 7  90 Σομαλία *
```

Προσπάθειες απελευθέρωση ναυτικών ισπανικού αλιευτικού Σομαλία