Groupe de travail Réseau
Request for Comments : 1439
Catégorie : Information

C. Finseth, University of Minnesota
mars 1993
Traduction Claude Brière de L'Isle

Unicité des identifiants uniques

Statut de ce mémoire
Le présent mémoire apporte des informations pour la communauté de l'Internet. Le présent mémoire ne spécifie aucune sorte de norme de l'Internet. La distribution du présent mémoire n'est soumise à aucune restriction.

Résumé
Cette RFC apporte des informations qui peuvent être utiles lors du choix d'une méthode pour allouer des identifiants uniques à des objets ou personnes.

1. Le problème

Les systèmes informatiques ont besoin d'un moyen d'identifier les personnes qui leur sont associées. Ces identifiants ont été appelés "nom d'utilisateur" ou "nom de compte". Les identifiants sont normalement de courtes chaînes alphanumériques. En général, ces identifiants doivent être uniques.

L'unicité est généralement réalisé par un des trois moyens suivants :

1) Les identifiants sont alloués de façon unique sans utiliser les informations associées à l'individu. Des exemples d'identifiants sont :

   ax54tv
   cs00034

   Cette méthode a souvent été utilisée par de gros systèmes en temps partagé. Bien qu'elle réalise la propriété d'unicité, il n'y a pas de moyen de deviner l'identifiant sans le connaître par d'autres moyens.

2) Les identifiants sont alloués d'une manière unique avec le corps de l'identifiant déduit algorithmiquement du nom de l'individu. Des exemples d'identifant sont :

   Craig.A.Finseth-1
   Finseth1
   caf-1
   fins0001

3) Les identifiants ne sont en général pas alloués de manière unique : l'identifiant est déduit par un algorithme du nom de l'individu et les dupliqués sont traités de façon ad-hoc. Des exemples d'identifiants sont :

   Craig.Finseth
   caf

Maintenant que la messagerie électronique est largement répandue, une caractéristique importante d'un système d'identifiant est la capacité à prédire l'identifiant sur la base des autres informations associées à l'individu. Ces autres informations sont normalement le nom de la personne.

Les méthodes deux et trois rendent de telles prédictions possibles, en particulier si on a un exemple de transposition du nom d'une personne en l'identifiant. La méthode deux s'appuie sur l'utilisation d'une partie ou de tout le nom et en le faisant varier algorithmiquement pour assurer l'unicité (par exemple, en ajoutant un entier). La méthode trois s'appuie sur l'utilisation d'une partie ou de tout le nom et en choisissant un identifiant de remplacement en cas de duplication.

Pour les deux méthodes, il est important de minimiser le besoin de faire les ajustements requis pour assurer l'unicité (c'est-à-dire, un entier qui ne soit ni 1 ni un identifiant de remplacement). La probabilité qu'un ajustement soit requis dépend du format de l'identifiant et de la taille de l'organisation.

2. Formats d'identifiant

Il y a un certain nombre de formats d'identifiants populaires. Cette section va faire une liste de certains d'entre eux et fournir les valeurs typiques et maximum pour le nombre d'identifiants possibles. Une valeur "typique" est le nombre qu'on va probablement rencontrer dans la vraie vie. Une valeur "maximum" est le plus grand nombre de valeurs possibles (sans aller jusqu'aux extrêmes). Toutes les gammes sont exprimées en nombre de bits.

2.1 Initiales

Il y a trois formats populaires fondés sur les initiales : ceux avec une, deux, ou trois lettres. (Le nombre de gens qui ont plus de trois initiales est supposé faible.) Les valeurs :

format	typique	maximum
I	4	5
II	8	10
III	12	15

On peut aussi les voir comme première (F), moyenne (M) et dernière (L) initiale :

I	4	5
F L	8	10
F M L	12	15

2.2 Noms

Là encore, il y a trois formats populaires fondés sur l'utilisation des noms : ceux avec le prénom, le nom de famille, et le prénom et le nom. Les valeurs :

format	typique	maximum
prénom	8	14
nom	9	13
prénom nom	17	27

2.3 Combinaisons

On peut voir ces combinaisons utilisées ("F" est première initiale, "M" est initiale moyenne, et "L" est la dernière initiale) :

format	typique	maximum
F nom	13	18
F M nom	17	23
prénom L	12	19
prénom M nom	21	32

2.4 Liste complète

Ici on a toutes les combinaisons possibles de rien, initiale, et nom complet pour first, middle, et last. Le nombre de noms moyens est supposé être le même que le nombre de premiers prénoms. Les valeurs :

format	typique	maximum
_ _ _	0	0
_ _ L	4	5
_ _ Last	9	13
_ M _	4	5
_ M L	5	10
_ M Last	13	18

_ Middle _	8	14
_ Middle L	12	19
_ Middle Last	17	27
F _ _	4	5
F _ L	5	10
F _ Last	13	18
F M _	5	10
F M L	12	15
F M Last	17	23
F Middle _	12	19
F Middle L	16	24
F Middle Last	21	32
First _ _	8	14
First _ L	12	19
First _ Last	17	27
First M _	12	19
First M L	16	24
First M Last	21	32
First Middle _	16	28
First Middle L	20	33
First Middle Last	26	40

3. Probabilités de dupliqués

Comme on peut le voir, le contenu des informations de ces identifiants n'excède en aucun cas 40 bits et le contenu normal des informations n'excède jamais 26 bits. Le contenu de la plupart d'entre elles est dans la gamme de 8 à 20 bits. Les dupliqués sont donc non seulement possibles mais probables.

La méthode utilisée pour calculer la probabilité de dupliqués est la même que celle du problème bien connu de "l'anniversaire". Pour un univers de N éléments, la probabilité de dupliqués avec X membres est exprimée par :

$$1 - \frac{N}{N} \times \frac{N-1}{N} \times \frac{N-2}{N} \times \ldots \times \frac{N-(X-1)}{N}$$

Un programme pour calculer cette fonction pour des valeurs choisies de N est donnée dans l'appendice, avec son résultat complet.

La colonne "1%" est le nombre d'éléments (personnes) avant qu'une organisation de cette taille (univers) ait 1 % de chances d'un dupliqué. De même pour 2 %, 5 %, 10 %, et 20 %.

bits	**univers**	**1 %**	**2 %**	**5 %**	**10 %**	**20 %**
6	64	2	3	4	5	6
7	128	3	3	5	6	8
8	256	3	4	6	8	12
9	512	4	6	8	11	16
10	1 024	6	7	11	16	22
11	2 048	7	10	15	22	31
12	4 096	10	14	21	30	44
13	8 192	14	19	30	43	61
14	16 384	19	27	42	60	86
15	32 768	27	37	59	84	122

16	65 536	37	52	83	118	172
17	131 072	52	74	117	167	243
18	262 144	74	104	165	236	343
19	524 288	104	147	233	333	485
20	1 048, 76	146	207	329	471	685
21	2 097 152	206	292	465	666	968
22	4 194 304	291	413	657	941	1369
23	8 388 608	412	583	929	1330	1936
24	16 777 216	582	824	1313	1881	2737
25	33 554 432	822	1165	1856	2660	3871
26	67 108 864	1162	1648	2625	3761	5474
27	134 217 728	1644	2330	3712	5319	7740
28	268 435 456	2324	3294	5249	7522	10946
29	536 870 912	3286	4659	7422	10637	15480
30	1 073 741 824	4647	6588	10496	15043	21891
31	2 147 483 648	6571	9316	14844	21273	30959

Par exemple, supposons une organisation où on choisit la forme "Premier Dernier". Cette forme a 17 bits (normal) et 27 bits (maximum) d'informations. La ligne pertinente est :

17	131 072	52	74	117	167	243

Pour une organisation de 100 personnes, la probabilité d'un dupliqué serait entre 2 % et 5 % (probablement autour de 4 %). Si l'organisation a 1 000 personnes, la probabilité d'un dupliqué serait très supérieure à 20 %.

Appendice. Réutilisation des identifiants et questions de confidentialité

Disons qu'une organisation choisit le format First.M.Last-# comme le fait ma propre organisation. Le -# est-il requis, ou peut on simplement faire :

Craig.A.Finseth

pour le premier et

Craig.A.Finseth-2

(ou -1) pour le second ? La réponse est "non", bien que pour des raisons non évidentes.

Supposons que l'organisation ait fait ce choix et qu'un tiers veuille envoyer un message électronique à Craig.A.Finseth. À cause de la Loi sur la confidentialité des communications électroniques de 1987, une organisation doit traiter avec soin la mesagerie électronique. Dans ce cas, il n'y a pas moyen pour l'utilisateur tiers de savoir fiablement que l'envoi à Craig.A.Finseth est (peut être) le mauvais parti. Par ailleurs, si le suffixe -# est toujours présent et que les tentatives d'envoyer le message à la forme sans suffixe sont rejetées, l'utilisateur tiers va réaliser qu'elles doivent avoir le suffixe afin d'avoir un identifiant univoque.

Pour des raisons similaires, les identifiants de cette forme ne devraient pas être réutilisés pendant la durée de vie du système de messagerie.

Appendice. Programme Perl pour calculer les probabilités

```
#!/usr/local/bin/perl

for $bits (6..31) {
      &Compute($bits);
      }
exit(0);
```

```
# ---------------------------------------------------------

sub Compute {
        $bits = $_[0];
        $num = 1 << $bits;
        $cnt = $num;

        print "bits $bitsnumber $num:0;

        for ($prob = 1; $prob > 0.99; ) {
                $prob *= $cnt / $num;
                $cnt--;
                }

        print "", $num - $cnt, "$prob0;

        for (; $prob > 0.98; ) {

                $prob *= $cnt / $num;
                $cnt--;
                }
        print "", $num - $cnt, "$prob0;

        for (; $prob > 0.95; ) {
                $prob *= $cnt / $num;
                $cnt--;
                }
        print "", $num - $cnt, "$prob0;

        for (; $prob > 0.90; ) {
                $prob *= $cnt / $num;
                $cnt--;
                }
        print "", $num - $cnt, "$prob0;

        for (; $prob > 0.80; ) {
                $prob *= $cnt / $num;
                $cnt--;
                }
        print "", $num - $cnt, "$prob0;

        print "0;
        }
```

Appendice. Résultat du programme Perl

```
bits 6          nombre 64 :
     2          0,984375
     3          0,95361328125
     4          0,90891265869140625
     5          0,85210561752319335938
     6          0,78553486615419387817

bits 7          nombre 128 :
     3          0,9766845703125
     3          0,9766845703125
     5          0,92398747801780700684
     6          0,88789421715773642063
     8          0,79999355674331695809
```

bits 8	nombre 256 :
3	0,988311767578125
4	0,97672998905181884766
6	0,94268989971169503406
8	0,89542306910786462204
12	0,76969425214152431547
bits 9	nombre 512 :
4	0,98832316696643829346
6	0,97102570187075798458
8	0,94652632751096643648
11	0,89748056780293572476
16	0,78916761796439427457
bits 10	nombre 1 024 :
6	0,98543241551841020964
7	0,97965839745873206645
11	0,94753115178840541244
16	0,88888866335604777014
22	0,79677613655632184564
bits 11	nombre 2 048 :
7	0,98978773152834598203
10	0,97823367137821537476
15	0,94990722378677450166
22	0,89298119682681720288
31	0,79597589885472519455
bits 12	nombre 4 096 :
10	0,98906539062491305447
14	0,97800426773009718762
21	0,94994111694430838355
30	0,89901365764115603874
44	0,79312138620093930452
bits 13	nombre 8 192 :
14	0,98894703242829806733
19	0,97932692503837115439
30	0,94822407309193512681
43	0,89545741661906652631
61	0,7993625840767998314
bits 14	nombre 16 384 :
19	0,98961337517641645434
27	0,97879319536756481668
42	0,94876352395820107155
60	0,89748107890372830209
86	0,79973683158771624591
bits 15	nombre 32 768 :
27	0,98934263776790121181
37	0,97987304880641035165
59	0,94909471808051404373
84	0,89899774209805793923
122	0,79809378598190949816
bits 16	nombre 65 536 :
37	0,98988724065590050216
52	0,97996496661944154649
83	0,94937874420413270737

118	0,89996948010355670711
172	0,79884228150816105618
bits 17	nombre 131 072 :
52	0,98993311138884398925
74	0,97960010416289267088
117	0,94952974978505377823
167	0,89960828942716541956
243	0,79894309171178368167
bits 18	nombre 262 144 :
74	0,98974844864797828503
104	0,97977315557223210174
165	0,94968621078621640041
236	0,8995926348279144058
343	0,7994422793765953994
bits 19	nombre 524 288 :
104	0,98983557888923057178
147	0,97973841652874515962
233	0,94974719445364064185
333	0,89991342619657743729
485	0,79936749144148444568
bits 20	nombre 1 048 576 :
146	0,98995567500195758015
207	0,97987072919607220989
329	0,94983990872655321702
471	0,89980857451706741656
685	0,79974215234216872172
bits 21	nombre 2 097 152 :
206	0,98998177463778547214
292	0,97994400939715686771
465	0,94985589918092261374
666	0,89978055267663470396
968	0,79994886751736571373
bits 22	nombre 4 194 304 :
291	0,98990013137747737812
413	0,97991951242142538714
657	0,94991674892578203959
941	0,89991652739633254399
1369	0,79989205747440361716
bits 23	nombre 8 388 608 :
412	0,98995762604049764022
583	0,97997846530691334888
929	0,94991024716640248826
1330	0,89999961063320443877
1936	0,79987028265451087794
bits 24	nombre 16 777 216 :
582	0,98997307486745211857
824	0,97999203469417239809
1313	0,94995516684099989835
1881	0,89997049960675035152
2737	0,79996700222056416063
bits 25	nombre 33 554 432 :
822	0,98999408609360783906

1165	0,9799956928177964155
1856	0,9499899669674316538
2660	0,8999664414095410736
3871	0,79992328289672998132
bits 26	nombre 67 108 864 :
1162	0,98999884535478044345
1648	0,9799801637652703068
2625	0,94997437525354821997
3761	0,89999748465616635773
5474	0,79993922903192515861
bits 27	nombre 134 217 728:
1644	0,9899880636014986024
2330	0,97998730103356856969
3712	0,94997727934463771504
5319	0,89998552434244594167
7740	0,79999591580103557309
bits 28	nombre 268 435 456:
2324	0,98999458855588851058
3294	0,97999828329325222587
5249	0,94998397932368705554
7522	0,89998576049206902017
10946	0,79999058777500076101
bits 29	nombre 536 870 912 :
3286	0,98999717306002099626
4659	0,97999160965267329004
7422	0,94999720388831232487
10637	0,89999506567702891591
15480	0,7999860979665908145
bits 30	nombre 1 073 741 824 :
4647	0,98999674474047760775
6588	0,97999531736215383937
10496	0,94999806770951356061
15043	0,89999250738244507275
21891	0,79999995570982085358
bits 31	nombre 2 147 483 648 :
6571	0,98999869761078929109
9316	0,97999801528523688976
14844	0,94999403283519279206
21273	0,89999983631135749285
30959	0,79999272222201334159

Références

Bruce Lansky (1984). "The Best Baby Name Book". Deephaven, MN: Meadowbrook. ISBN 0-671-54463-2.

Lareina Rule (1988). "Name Your Baby". Bantam. ISBN 0-553-27145-8.

Considérations sur la sécurité

Les questions de sécurité ne sont pas discutées dans le présent mémoire.

Adresse de l'auteur

Craig A. Finseth
Networking Services
Computer and Information Services
University of Minnesota
130 Lind Hall
207 Church St. SE
Minneapolis, MN 55455-0134
USA
mél : Craig.A.Finseth-1@umn.edu
téléphone : +1 612 624 3375
Fax : +1 612 626 1002