UN ALGORITHME QUI COMPTE EN TIRANT AU SORT

PHILIPPE FLAJOLET ET FRÉDÉRIC MEUNIER

Plusieurs chercheurs du Projet Algorithmes de l'Inria Rocquencourt — Philippe Flajolet, Eric Fusy et Frédéric Meunier, en collaboration avec Olivier Gandouet du Laboratoire d'Informatique LIRMM de Montpellier — ont récemment mis au point le meilleur algorithme connu d'estimation de cardinalité de grands ensembles. Il s'agit d'évaluer le nombre d'objets différents dans un ensemble pouvant en contenir des milliards. Dans les applications informatiques, il n'est pas envisageable de construire une liste complète de ces objets. L'algorithme HyperLogLog parvient à déterminer ce nombre avec une précision de 2% en utilisant une mémoire équivalente à 1500 caractères (octets). C'est un peu comme si, assistant à une pièce de théâtre, on arrivait à estimer précisément le nombre de mots différents prononcés au cours de la pièce, en disposant en tout et pour tout d'un crayon, d'une gomme et d'un quart de feuille A4.

Une application de cet algorithme est la détection d'attaques sur internet. En effet, une attaque se caractérise souvent par une augmentation du nombre de connexions différentes au niveau d'un routeur, nombre qu'HyperLogLog estime parfaitement en ligne. Une autre application est la mesure de similarité dans de grandes bases de documents. En ce cas, HyperLogLog associe à chaque document une signature, dont on tire les informations de cardinalité. C'est cette signature qui permet d'estimer la proportion d'éléments communs à deux documents, laquelle constitue un bon indice de similarité et est exploitable par les moteurs de recherche.

L'algorithme HyperLogLog s'inspire des travaux d'une ancienne doctorante du Projet Algorithmes, Marianne Durand-Maurel, lors de sa thèse en 2004. Le point de départ est une "randomisation" du problème: chaque élément déclenche un tirage pseudo-aléatoire d'un nombre réel, différentes occurences de l'élément donnant lieu au même tirage. C'est à partir d'observations numériques précises sur cette suite de tirages qu'est calculée une estimation de la cardinalité. La mise au point de la formule permettant ce calcul a nécessité des techniques mathématiques de haut vol, comme le moyennage stochastique, la dépoissonisation analytique et la transformée de Mellin. L'informatique doit souvent s'appuyer sur des méthodes mathématiques poussées et il n'est pas rare, comme c'est le cas ici, qu'elles occupent une place incontournable. Pas de mathématiques, pas de formule; pas de formule, pas d'algorithme !