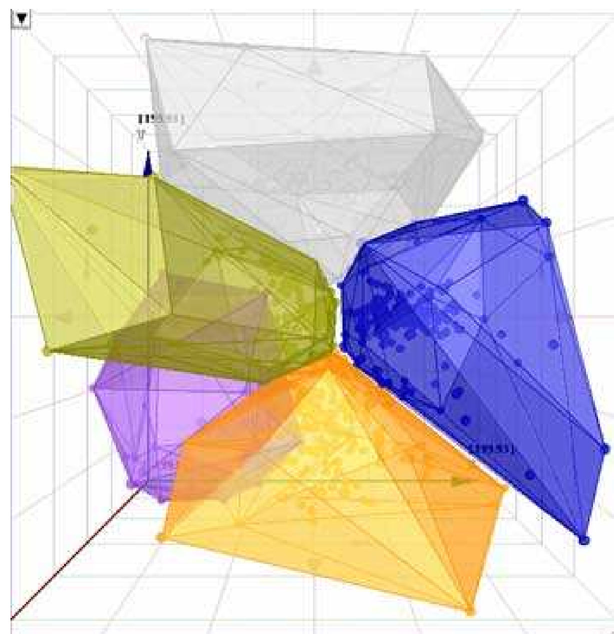
Géométrie algorithmique en très grandes dimensions (Proposition de stage X2005 — 2007)

Frank NIELSEN
(E-mail :Frank.Nielsen@acm.org)

Thématique	géométrie algorithmique, algorithmique
Laboratoire	LIX, École Polytechnique, Paris
Durée	4 à 5 mois (Avril 2008 $\sim$, anglais ou français)

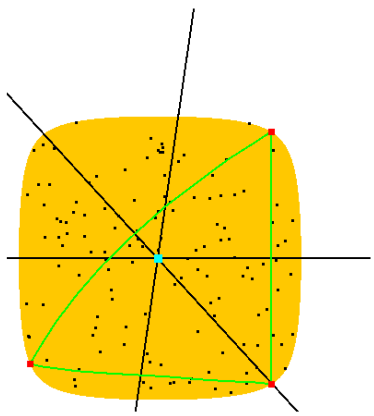
Positionnement du stage



La géométrie algorithmique s'est traditionellement intéressée aux “petites” dimensions, c'est-à-dire au cas où le nombre de données en entrée n est largement supérieur à la dimension de l'espace ambiant d : $n >> d$. Dans ce stage, nous nous intéressons au cas contraire ($d >> n$, avec disons $d > 10000$) pour un certain nombre de problèmes fondamentaux [1,2] comme les structures de données pour la localisation des plus proches voisins d'un point requête donné, le clustering [3,4], ou bien encore la boule englobante (quasi-)minimale [5], etc. Bien que théoriquement, nous savons qu'il existe un espace affine de dimension au plus $d' = n - 1$ contenant l'ensemble des données, nous ne pouvons pas projeter directement les données dans cet espace (une réecriture compacte des coordonnées sur d' dimensions), car cela ferait appel à des calculs de déterminant qui auraient pour effet de détériorer désastreusement la précision numérique des entrées, et de les rendre complétement inutilisables si l'on utilise une arithmétique fixe traditionnelle (sur 32 ou 64 bits). Le stage se propose donc d'analyser en *amont* les récents résultats [1,2,3,4,5] en géométrie algorithmique portant sur les grandes dimensions et de réflechir sur des problèmes d'actualité connexes. Notamment, on fera une classification des algorithmes en deux sous-classes : les algorithmes facilement implantables et les autres, restant pour l'instant purement théoriques. En aval, ces travaux ont des débouchés importants en infographie (par exemple pour l'analyse/synthèse de texture) et en apprentissage par ordinateur (support vector et core-vector machines).

Objectifs du stage

Le stage consiste en deux étapes. Tout d'abord, on s'attaquera à la rédaction d'un état de l'art *concis* en anglais sur les principaux paradigmes permettant de traiter des problèmes en grandes dimensions (eg., core-sets, réduction de dimension, apprentissage de variétés, plongement de métriques, randomisation, etc.), en classant les méthodes suivant leurs degrés de complexité vis à vis d'une implantation. On analysera notamment les différentes représentations de données : brutes, représentation succintes, représentation avec mémoire auxiliaire, représentation par flot (data streams), etc.



La seconde partie consistera à implanter deux algorithmes fondamentaux en très grandes dimensions qui utilisent comme brique de base des requêtes de plus proches voisins : le clustering par la méthode des centroïdes/médioïdes (k-means), et une approximation arbitrairement fine de la plus petite boule englobante (problème d'optimisation de type MiniMax) pour des distances non-Euclidiennes [6]: les divergences de Bregman[1], Csiszár ou encore de Burbea-Rao qui ont des significations axiomatiques précises et distinctes en théorie de l'information. Dans ce cadre, on regardera les implantations de bibliothèques existantes portant sur les core vector machines (CVMs) (`http://www.cse.ust.hk/~ivor/cvm.html`) et les core-sets. En guise d'applications, on considérera le calcul du "centre" d'un ensemble d'histogrammes (représenté par un ensemble de points en grandes dimensions) provenant d'images couleurs de textures prises sous des conditions d'éclairage variables (`www.cs.columbia.edu/CAVE/software/curet/`).

Profil/prérequis

Mots clefs: clustering, minimax, distances.
Outils: Java ou C++ (au choix).

Bibliographie

1. Alexandr Andoni, Piotr Indyk: Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions. FOCS 2006: 459-468

2. Sudipto Guha, Piotr Indyk, Andrew McGregor: Sketching Information Divergences. COLT 2007: 424-438

3. David Arthur, Sergei Vassilvitskii: k-means++: the advantages of careful seeding. SODA 2007: 1027-1035

4. Sariel Har-Peled, Akash Kushal: Smaller Coresets for k-Median and k-Means Clustering. Discrete & Computational Geometry 37(1): 3-19 (2007)

5. Frank Nielsen, Richard Nock: On approximating the smallest enclosing Bregman Balls. Symposium on Computational Geometry 2006: 485-486

6. Elena Deza, Michel Marie Deza: Dictionary of distances, ISBN 0-444-52087-2, Elsevier, 2005.

[1] `http://www.sonycsl.co.jp/person/nielsen/BregmanBall/BBC/`