

Recherche de plus proches voisins inverses en grandes dimensions

Steve Y. Oudot — INRIA Saclay, équipe Geometrica

Sujet de stage

Thématique : algorithmique

Laboratoire, institution ou université : INRIA Saclay – Île-de-France

Ville et pays : Orsay, France

Équipe ou projet dans le laboratoire : Geometrica

Directeur de stage : Steve Oudot (steve.oudot@inria.fr)

Présentation générale du domaine : La recherche de plus proches voisins est un problème central en informatique, où elle trouve des applications dans des domaines aussi variés que l'apprentissage, les bases de données, ou la géométrie. Étant donné un nuage de points P , le but est de pré-calculer une structure de données permettant de retrouver, pour tout point de requête q , le plus proche voisin de q dans P . De nombreuses méthodes ont été développées pour répondre à ce problème, dont la quasi totalité souffre de ce que l'on appelle le "fléau de la dimension", à savoir qu'elles requièrent soit un espace mémoire de taille exponentiel en la dimension des données, soit un temps de calcul exponentiel. Le résultat en grandes dimensions est qu'elles sont soit inutilisables, soit à peine meilleures qu'une simple recherche exhaustive sur les points du nuage.

A la toute fin des années 90 est apparue une nouvelle approche, appelée Locality-Sensitive Hashing (LSH), qui permet de répondre de manière approchée à des requêtes de plus proches voisins en temps sous-linéaire et en espace polynômial. L'introduction de cette technique a révolutionné le domaine, en démontrant que le fléau de la dimension n'était peut-être pas une fatalité. Depuis, elle a connu un essor très important, tant du point de vue théorique que pratique, avec l'apparition de bibliothèques C++ comme LSHKIT [3].

Objectifs du stage : Le stage se focalisera sur le problème inverse, à savoir : étant donné un nuage de points P , construire une structure de données qui permette de retrouver, pour tout point de requête q , les points de P ayant q comme plus proche voisin. Bien que fondamental et fortement lié à la recherche de plus proches voisins classique, ce problème n'a été abordé sous un angle théorique que depuis quelques années [2], et sa complexité en grandes dimensions reste encore mal connue. Récemment, D. Arthur, S. Oudot et A. Sharma [1] ont montré comment la technique de LSH peut être utilisée pour le résoudre en temps sous-linéaire et en espace polynômial, une première dans ce domaine.

L'objectif principal du stage sera d'implémenter et de tester l'algorithme de Arthur, Oudot et Sharma sur des données réelles, afin d'en mieux cerner les atouts et les limites. En particulier, on tentera de déterminer si ses qualités théoriques (sous-linéarité, faible espace mémoire) se

vérifient en pratique, ou bien si son comportement est lui aussi soumis au fléau de la dimension. On le comparera également avec d'autres techniques comme celle de Korn et Muthukrishnan [2] basée sur l'utilisation de R-trees. Pour cela, l'étudiant devra d'abord se familiariser avec la technique du LSH. L'implémentation se fera de préférence en C++, mais si besoin l'utilisation d'un autre langage de programmation sera possible. Une particularité de l'algorithme est d'utiliser la technique de LSH comme une boîte noire, ce qui permettra de s'appuyer sur des bibliothèques existantes comme LSHKIT pour l'implémentation.

Compétences requises :

- une connaissance solide des structures de données classiques, et en particulier des tables de hachage,
- des notions d'analyse des algorithmes,
- quelques rudiments de théorie des probabilités (variables aléatoires, indépendance, inégalité de Boole, etc.),
- la maîtrise d'un langage de programmation, de préférence C++ mais pas obligatoirement (Java ou Caml feront très bien l'affaire),
- malgré la nature très géométrique du problème, il ne sera pas nécessaire d'avoir suivi le cours de géométrie algorithmique.

Note : durant ce stage, l'étudiant aura l'occasion d'interagir non seulement avec les membres de l'équipe Geometrica à l'INRIA, mais également avec des chercheurs et étudiants du groupe d'informatique théorique à l'université de Stanford, dont David Arthur et Aneesh Sharma.

Références

- [1] D. Arthur, S. Oudot, and A. Sharma. Finding friends and followers in sub-linear time. Research Report 7084, INRIA, November 2009.
- [2] Flip Korn and S. Muthukrishnan. Influence Sets Based on Reverse Nearest Neighbor Queries. *ACM SIGMOD Record*, 29(2) :201–212, 2000.
- [3] LSHKIT : A C++ Locality-Sensitive Hashing Library (<http://lshkit.sourceforge.net/>).