k-NN et Hadoop Map-Reduce

Fabrice Huet fabrice.huet@unice.fr
Justine Rochas Justine.rochas@inria.fr

11 février 2014

Nombre d'étudiants souhaités : 3-4

Description du sujet

Les algorithmes de type k-NN (k Nearest Neighbors) sont des algorithmes très utilisés en intelligence artificielle en traitement d'images. Étant donné un ensemble de données N et une donnée de référence R, il fournit les k données de N qui sont les plus proches de R. La proximité est définie par une distance (au sens mathématique du terme) qui dépend du problème et peut être par exemple une ressemblance entre des images.

Une façon simple d'implémenter un k-NN est de calculer la distance de R avec tous les éléments et ensuite prendre les k plus petits. Intuitivement, on sent bien que cette façon naïve n'est pas la plus efficace et effectivement, il en existe plusieurs variantes nécessitant un prétraitement des données.

Quand le nombre de données est important, une seule machine ne suffit pas et on a recours à de la programmation distribuée, comme par exemple Hadoop Map-Reduce. Nous avons développé une variante de Hadoop, Continuous Hadoop, qui permet de sauvegarder des résultats intermédiaires afin de les réutiliser par la suite. Le but de ce TER est d'implémenter plusieurs variantes de k-NN en Hadoop et en Continuous Hadoop et comparer leurs avantages et inconvénients.

Travail demandé :

1. Implémenter deux variantes de k-NN en Hadoop et Continuous Hadoop
2. Proposer et effectuer des expérimentations permettant de comparer ces variantes

Les expérimentations seront effectuées sur Grid'5000, un Cloud national de plusieurs centaines de machines.

Lieu

INRIA Sophia Antipolis

Prérequis

Bonne connaissance de Java, connaissance des algorithmes parallèles et distribués. Connaissance du modèle MapReduce est un plus.

Informations complémentaires