II. Ma mission

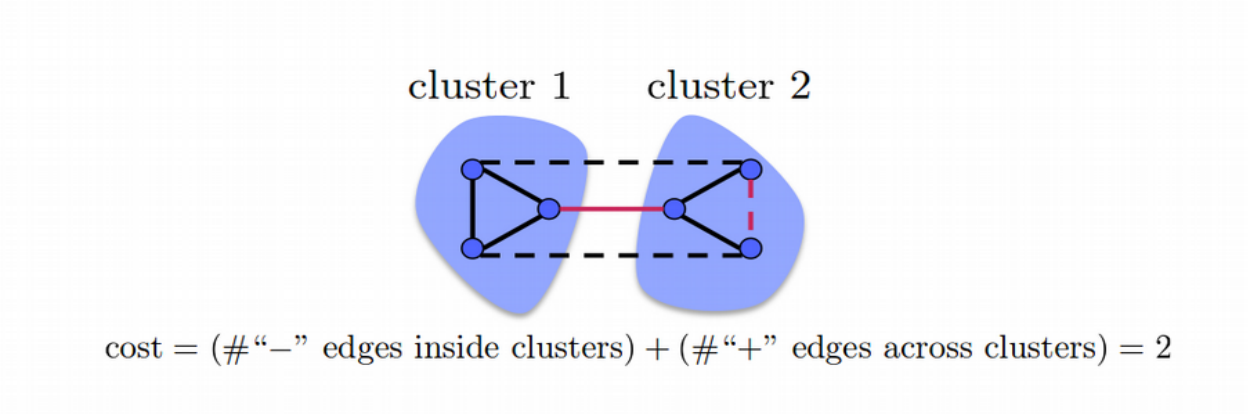
1. Ma mission et ses objectifs

L'équipe MAGNET axe son projet sur le Machine Learning, qui est un champ d'étude tendant à rendre possible pour un ordinateur, d'effectuer des tâches pour lesquelles il n'est pas explicitement programmé, en apprenant à travers des données. Cette discipline qui vise à faire des prédictions, se base sur des jeux de données, qui peuvent être de nature différentes : des structure de données continues comme une courbe, ou discrètes comme un arbre ou un graphe. Cette dernière structure, les graphes, est celle sur laquelle j'ai travaillé durant ces deux mois. En effet l'objectif technique de mon stage est de réaliser un état de l'art et ~~faire une implémentation~~d'implémenter des algorithmes performants de « Correlation Clustering ».

Le Clustering est un problème de partitionnement de données en fonction de leur similarité, et plus spécifiquement le « Correlation Clustering » donne une méthode de clustering pour un ensemble afin d'obtenir un nombre optimal de clusters (i.e regroupement de points similaires) sans le spécifier à l'avance. Il intervient quand on opère sur un ensemble de données dont on ne connais pas ~~leurs~~de représentation explicite, mais seulement leurs relations (ex : les liens entre pages internet représente~~nt~~ leurs relations, bien qu'on ne connaisse pas le contenu de ces pages). L'objectif est alors de minimiser le nombre d'arêtes positives entre deux clusters (points se ressemblant, rassemblés dans des clusters différents), plus le nombre d'arêtes négatives à l'intérieur des clusters (points ne se ressemblants pas, et rassembl~~er~~és dans le même cluster). Ce problème peut être généralis~~er~~é par la formule suivante :

$$\mathrm{OPT} = \min_{1 \leq k \leq n} \min_{\substack{C_i \cap C_j = 0, \forall i \neq j \\ \cup_{i=1}^{k} C_i = \{1,\dots,n\}}} \sum_{i=1}^{k} E^{-}(\mathcal{C}_i, \mathcal{C}_i) + \sum_{i=1}^{k} \sum_{j=i+1}^{k} E^{+}(\mathcal{C}_i, \mathcal{C}_j)$$

ou E^- représente l'ensemble des arêtes négatives, tandis que E^+ représente l'ensemble des arêtes positives. Enfin Ci représente le i-ème cluster.



Exemple de correlation clustering, extrait de la publication « Scaling up Correlation Clustering
through Parallelism and Concurrency Control », écrit par Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran et Michael I. Jordan.

Une des limites du correlation clustering est le temps d'exécution de ces algorithmes, dû à la quantité de données qu'ils reçoivent en entrée. L'une des solutions à ce problème est de programmer ces algorithmes pour qu'ils utilisent la technologie dite « Multi-Thread », c'est à dire exploiter le caractère multi-core des processeurs, en divisant la tâches en sous problèmes, chacun ~~exécuté~~résolu ~~sur~~par une partie indépendante du processeur.

Ainsi la mission qui m'a été confié~~,~~ est l'implémentation d'un algorithme de correlation clustering en parallèle (i.e en utilisant le multi-threading) ainsi que son homologue en série (sans le multi-threading) afin de comparer les résultats en temps et en précision. La finalité du stage est d'exécuter ces programmes sur un des clusters de calcul de l'INRIA, afin de garantir des résultats en temps acceptable pour de gros graphes. Pour cela j'ai dû acquérir des connaissances théoriques sur le multi-threading, et tous ce qui permet son utilisation (synchronisation, lock, event...) ainsi que sur la théorie des graphes. Et enfin des connaissances pratiques sur le langage python que j'avais peu pratiqué avant le début de ce stage, le langage java, que je n'avais jamais utilis~~é~~é, ainsi que l'utilisation des systèmes Linux, et ses environnements de programmation.

L'intérêt de ma mission est de fournir à l'ensemble de l'équipe une ~~expérience témoin~~ implémentation de référence de correlation clustering, afin qu'ils puissent comparer en termes de précision~~s~~ et de temps d'exécution~~s~~ les résultats de leurs recherches~~, et~~ avec les approches existantes, ~~ceux déjà fournis~~ que j'ai implémenté~~e~~s à travers un état de l'art.

La recherche d'information constitue la première partie de mon stage. En effet, nous avions convenu, mon responsable et moi, après quelque recherche que j'utiliserai le langage python. Cette décision étai~~s~~t basée sur deux constats : le python est ~~très commode~~ aisé à manipuler car c'est un langage qui offre des outils de haut niveau ainsi qu'une syntaxe simple à utiliser, et est utilisable sur différents systèmes, ce qui permet de l'exécuter sur un cluster de calcul. Cependant il comporte un défaut majeur concernant le multi-thread : en tant que langage interprété, il comporte un « Global Interpreter Lock (GIL) » qui oblige les thread (fils d'exécution) à s'exécuter séquentiellement. Ainsi

l'intérêt de la technologie multi-thread est perdu. Toutefois, grâce au fait que python soit placé en licence libre, différents interpréteurs sont disponibles pour la communauté, dont « pypy-stm (pour Software Transactional Memory) » qui est une alternatives à l'interpréteur par défaut, ~~sc~~ensé permettre le multi-thread. Toutefois ce système, encore expérimental, reste lent pour les nombreuses opérations qui influent sur les structures de mémoires.

Pour pallier ~~aux~~ les défauts de python, nous avons décidé de reprendre l'algorithme sous java. En effet, java présente une technologie multi-thread qui a fait~~s~~ ses preuves, mais ~~d'obliger ce défaut~~à ~~l~~contraint à employer une ~~l'~~approche orienté objet de la programmation.

2.Solution

La solution est donc de programmer cet algorithme en java, afin d'utiliser au mieux la technologie multi-thread. Après quelques recherches l'algorithme prend forme, s'appuyant sur la programmation orientée objet : chaque tache indépendante de l'algorithme est distribuée à un objet créé à cet effet. Ainsi un objet li~~st~~ le graphe et construit sa ~~structure~~représentation en mémoire, tandis qu'un autre crée les listes de travail nécessaires aux threads de clustering (des objets eux aussi), et ainsi de suite. Par ailleurs, après une implémentation efficace de l'algorithme de correlation clustering, nous avons pensé à un « post-processing », qui prend en entrée les résultats du premier algorithme et les ~~travaille~~modifie afin d'améliorer la précision (la précision est la minimisation de la fonction de coût présentée précédemment). En effet, le correlation clustering ~~n'étant pas déterministe (i.e ne renvoie pas les mêmes résultats pour les mêmes entrées)~~étant un problème dont la solution optimale ne peut être obtenue en temps polyniomal, l'algorithme n'en produit qu'une approximation, qui plus est non déterministe. Néanmois il est possible d'améliore a posteriori cette approximation.~~, le clustering optimal n'est jamais atteint.~~ La solution de post-processing est de calculer une matrice, dont les A(i,j) représente le gain en précision acquis grâce à la fusion des clusters i et j. Cette solution étant très coûteuse en opérations, je l'ai implément~~er~~ée encore une fois en utilisant la technologie multi-thread. Chaque thread traitant ainsi une ligne de la matrice. Enfin, la fusion apportant le plus grand gain est effectuée, et la matrice est recalculée, jusqu'à ce qu'aucune fusion n'apporte de gain (i.e tous les A(i,j) sont négatifs ou nuls).

3.Résultat

Au terme de ces deux mois, l'algorithme de correlation clustering est terminé, et le post-processing également. Afin de pouvoir visualiser les qualités que présente le multi-thread, j'ai créé un script shell (linux) qui a exécut~~er~~é de nombreuses fois cet algorithme, paramétré à 32 thread parallèles, et en a retenu les temps d'exécution, sur un des cluster de calcul de l'INRIA. Par ailleurs, un autre script avais le même rôle mais pour un seul thread cet fois (équivalent de la version série de l'algorithme).

Les test se sont faist sur deux graphes : un très massifs, mais pour lequel le post-processing était trop coûteux (500 000 arrêtes), et enfin un plus petit (100 000 arrêtes). Dans les deux cas le multi-thread se révèle très avantageux puisqu'il permet de diviser les temps d'exécution par 10 dans la plupart des cas, par 15 dans les meilleurs des cas.

Par ailleurs, après avoir étudié attentivement les résultats sur de petits graphes, je me suis rendu comptes que beaucoup d'imprécisions étaient dues aux « clusters singletons », qui sont des clusters ne comportant qu'un unique point. Mon idée est alors de créer une étape intermédiaire, avant les calculs des matrices de fusions, afin de traiter ces clusters singletons en particulier. Cette étape se révèle très péerformante puisqu'elle divise en moyenne par deux le nombre de fusions engendrées par le post-processing, et divise donc par deux le nombre de matrice à calculer.

Au final, au terme de ces deux mois de stages, je livre à l'équipe MAGNET de l'INRIA un algorithme de correlation clustering complémenté par le post-processing, capable en moins d'une minute de rassembler par similarité 100 000 points, avec une erreur de 0,01 % en moyenne.

3. Mon point de vue sur l'INRIA et sur mon stage

Tout d'abord, les dates de mon stages ne m'ont pas permis de voir le fonctionnement de l'INRIA dans sa globalité : en effet, il n'y a pas de conférences, ni d'invités étrangers pendant l'été, et le métier de chercheur durant cette période ne se résume finalement qu'à un travail de bureau. Cependant je reste persuaderé que l'INRIA, de par ses collaborations, conférences et autres activités offres une diversité intéressante de techniques et de sciences à ses employés. Par ailleurs, j'ai découvert pendant mon stage l'importance qu'accorde l'INRIA au transfert des connaissances et des compétences. Je trouve personnellement que cette dimension de la recherche est primordiale, et pas assez souvent mise en valeur. De plus je pense qu'elle s'inscrit bien dans le travail de l'ingénieur, puisqu'il est l'outil entre les connaissances et l'industrie.

Par ailleurs, j'apprécie le concept d'équipe-projet, aux objectifs précis et aux échéances fixées, car elle permettent une vision plus proche de l'application de la recherche. En effet, ce concept contre balance l'idée générale de la recherche en science pour la science...

De plus, à titre personnel je suis très intéressé par de nombreux projets de l'INRIA, et surtout par les thèmes que l'institut développe : intelligence artificielle, big data, internet des objets, bio-technologies etc...

Enfin, j'ai apprécier ce stage car il m'a permis d'acquérir beaucoup de connaissances techniques en informatique, et ceci en autonomie. Il m'a en outre permis de découvrir le monde de la recherche, ses enjeux et son fonctionnement. J'ai ainsi pu affiner mon projet d'étude, en tenant compte de ce que j'ai pu observer pendant deux mois.

5. Conclusion

En conclusion ce stage à rempli les deux objectifs que je m'étais fixés : le premier était de prendre connaissance du monde du travail, et plu spécifiquement le monde de la recherche. Le second était une mise en pratique ainsi qu'une acquisition de compétences liées à la programmation et à l'informatique en général.

Tout d'abord, j'ai pu découvrir le monde de la recherche dans un cadre adapté à la fois à mes études (recherches en informatiques), et adapté à moi-même, puisque j'ai travaillé dans une petite équipes-projet, ce qui m'a permis de créer des liens, notamment avec mon responsable, mais m'a aussi permis de comprendre l'enjeu du travail de chacun, dans une ambiance décontractée, studieuse et intimiste. J'y ai découvert des personnes passionnés et impliqués, ainsi qu'un monde du travail qui allie curiosité et rigueur intellectuelle. J'ai également pu apprendre les rouages de la recherche scientifique, ses démarches, comme le transfert de compétences et de connaissances, et ses impératifs, comme les publications de recherches, et enfin la nature collaborative de ce milieu. Puis j'ai découvert, grâce à la rédaction de ce rapport, l'histoire de l'INRIA, qui marque profondément le paysage scientifique français, et le place à un niveau de compétitivité internationale.

Enfin, ce stage m'a beaucoup apporté sur les connaissance : J'ai premièrement appris deux langages, sur des thématiques qui ne me semblaient pas acquises au commencement de ces deux mois. D'une part python, qui m'apparaît comme un langage très intuitifs, et facile à manier. De plus j'ai appris quantité d'informations sur ce langage, en dehors de l'aspect programmation : comment il a été pensé, créé, et comment une communauté de scientifiques continue de le mettre au point. Par ailleurs j'ai également appris le java, ainsi que beaucoup de chose qui l'entoure : ses bibliothèques, ses implémentation etc.. Finalement, j'ai aussi appris beaucoup de concepts théoriques, tel que le multi-thread, et tous ce qu'il implique, mais aussi l'utilisation de serveur à distance, ou même la répartition de la mémoire à l'intérieur d'un processeur multi-core.

Pour finir, ce stage m'apporte deux perspectives : la première, à court terme, concerne le PPE de ma 4ème année à l'ECE. De nombreuses connaissances que j'ai acquises me seront utiles au bon déroulement du projet, puisqu'il concerne essentiellement du développement de software. La seconde, sur le plus long terme, concerne mon projet de vie et de carrière. En effet j'ai découvert avec plaisir le monde de la recherche, et mis en comparaison avec un futur stage en entreprise, il me permettra de faire un choix entre la recherche, ou un poste d'ingénieur.