le cnam

Sujet UE RCP216
Ingénierie de la fouille et de la visualisation des données massives

Année universitaire 2018–2019

Examen 1ère session : 29 janvier 2019

Responsable : Michel CRUCIANU

Durée : 2h00

Seul document autorisé : 2 pages A4 recto-verso, manuscrites.

Les téléphones mobiles et autres équipements communicants (PC, tablette, etc.) doivent être éteints et rangés dans les sacs pendant toute la durée de l'épreuve. Calculatrice autorisée mais inutile.

Sujet de 6 pages, celle-ci comprise.

Vérifiez que vous disposez bien de la totalité des pages du sujet en début d'épreuve et signalez tout problème de reprographie le cas échéant.

---

1 Fouille de données (13 points)

[3 pts] Question 1 : Pourquoi le stockage des données intermédiaires est important dans un environnement massivement parallèle ? Pourquoi cela ralentit fortement le déroulement des algorithmes itératifs ? Comment est résolue cette contradiction dans Spark ?

Correction :
Dans un environnement massivement parallèle les pannes de certains nœuds sont inévitables, le stockage fiable des données intermédiaires permet de ne pas avoir à refaire les itérations précédentes pour obtenir des résultats intermédiaires perdus suite aux pannes. Le stockage sur un support non volatile est fait à un débit très inférieur au débit du stockage en mémoire vive (voir dans le cours introductif la figure avec la hiérarchie de stockage), ce stockage prolonge donc le temps nécessaire entre deux itérations successives. Spark préfère éviter le stockage sur support non volatile des données intermédiaires et les recalcule si elles sont perdues ; le calcul est limité aux données perdues.

[2 pts] Question 2 : La « malédiction de la dimension » se manifeste plus pour les représentations Word2Vec ou pour les représentations ESA (*Explicit Semantic Analysis*) ? Pourquoi ?

Correction :
Les représentations ESA sont de dimension beaucoup plus grande que les représentations Word2Vec, la « malédiction de la dimension » devrait donc se manifester plus pour les représentations ESA. Mais la distribution des données a aussi son importance.

[2 pts] Question 3 : Expliquez pourquoi un filtre de Bloom présente un comportement asymétrique : il garantit l'absence de faux négatifs mais ne garantit pas l'absence de faux positifs.

Correction :
L'absence de faux négatifs est garantie par construction : la liste exhaustive des vrais positifs est fournie à la construction du filtre et la valeur de *hash* de chaque vrai positif est associée à un bit à 1 qui garantit la détection. En revanche, les fonctions de hachage ne peuvent pas éviter les collisions (même *hash* pour deux données différentes), certains vrais négatifs entreront en collision avec de vrais positifs et deviendront donc des faux positifs.

[2 pts] Question 4 : Pour classer des données en flux il est possible d'employer *Streaming K-means* sur le flux. Il est également possible d'appliquer K-means sur la tranche initiale du flux pour trouver les centres, ensuite d'utiliser ces centres sur les tranches ultérieures

pour classer les données. Quels avantages et inconvénients présente chacune de ces méthodes ?

Correction :
Streaming K-means permet de faire évoluer les centres suivant l'évolution de la distribution des données et est ainsi mieux adapté à une distribution non stationnaire. L'autre solution est en revanche moins coûteuse car les centres de départ ne sont pas mis à jour pour chaque tranche du flux. Par ailleurs, même si la distribution est stationnaire, la tranche initiale peut contenir un échantillon de taille trop faible, peu représentatif de cette distribution ; tenir compte des données de plusieurs tranches permet d'obtenir de meilleurs résultats. En revanche, dans l'implémentation Spark de *Streaming K-means* l'initialisation est basique (tirage aléatoire simple), alors que l'implémentation de K-means permet d'utiliser une technique d'initialisation plus évoluée (K-means||).

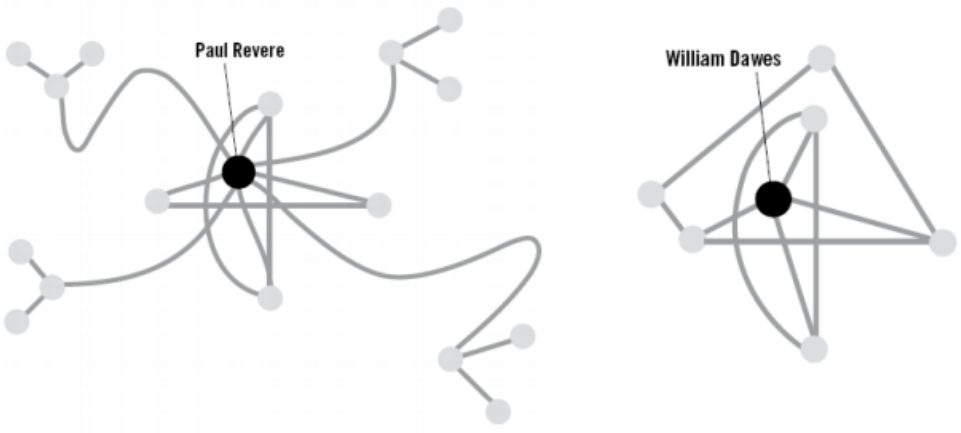
[1pt] Question 5 : Qu'est-ce qu'une distribution « en loi de puissance » ? Dans quel(s) contexte(s) en rencontre-t-on souvent pour la fouille de graphes/réseaux sociaux ?

Correction :
Voir support de cours.

[1pt] Question 6 : Expliquez en quoi calculer un indicateur simple comme la moyenne dans un cas de « loi de puissance » est peu pertinent.

Correction :
Voir la Fig. 102 dans le support de cours.

[2pt] Question 7 : En 1775, lors de l'arrivée des Anglais pour mater les révolutionnaires américains, Paul Revere et William Dawes ont chevauché de Boston jusqu'à Lexington pour avertir leurs camarades. Ils disposaient des réseaux de connaissances suivants :



Selon vous, qui a propagé le plus efficacement la nouvelle ? Expliquez pourquoi à partir des structures comparées de ces deux réseaux. (**1 point**)

Correction :
Paul Revere a un réseau plus grand, et moins clusterisé que Dawes. Il a pu toucher des « hubs » locaux, qui ont à leur tour averti d'autres personnes. La centralité est une caractéristique fondamentale ici, ainsi que la taille du réseau.

---

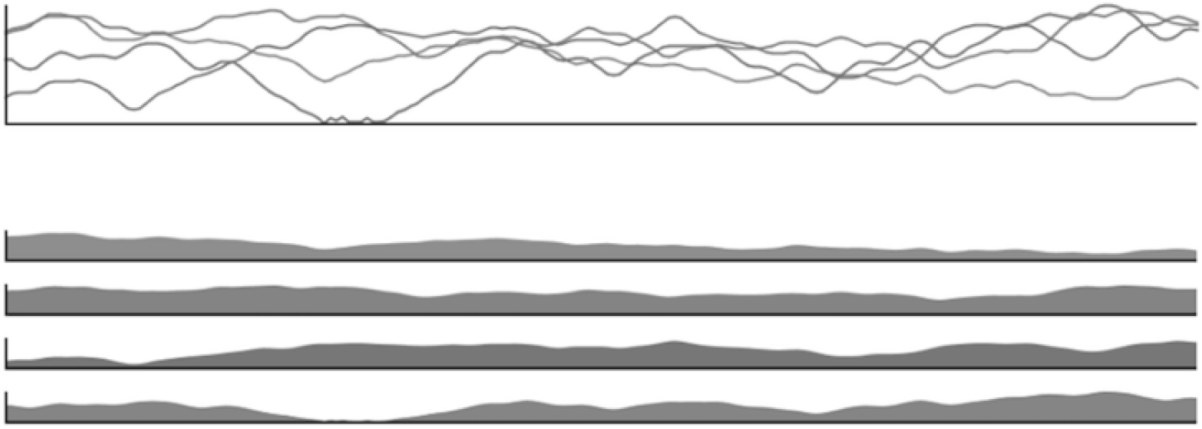
2 Visualisation et interaction (7 points)

[2 pts] Question 8 : Nous avons étudié en cours deux représentations spécifiques aux arbres : les Treemaps et les arbres hyperboliques. Donner deux avantages communs à ces techniques par rapport à la représentation classique des arbres sous forme de boites et flèches.

Correction :
Les deux représentations en commun 1) le fait qu'elles se basent sur un algorithme de dessin déterministe, de faible complexité en temps. 2) elles prennent comme donnée de départ la dimension de la zone de tracé, donc le tracé va s'adapter à cette surface quel que soit le nombre de sommets de l'arbre.

On peut aussi supposer que ces techniques facilitent la comparaison d'arbres, par rapport au dessin sous forme de boites et flèches.

[5 pts] Question 9 : La représentation des séries chronologiques par des courbes est une idée très ancienne et très efficace. Plusieurs méthodes ont été inventées pour le cas des séries multiples devant être étudiées simultanément. En 2010, des chercheurs ont étudiées l'efficacité des ces différentes méthodes. La figure ci-dessous, extraite de leur publication, montre deux techniques. La méthode 1, à haut, consiste à superposer les courbes des différentes séries en leur donnant une plage de variation commune. La méthode 2, en bas, représente les séries les unes sur les autres, en réduisant au maximum la hauteur attribuée aux plages de variation. Attention, la figure est ici reproduite en niveaux de gris.



1. Comment s'appelle la technique employée pour la méthode 2 ? Elle s'applique aussi à d'autres visualisations que des courbes. **[1 pt]**
2. Pour que la méthode 1 fonctionne bien, il faut dessiner chaque courbe avec une couleur différente. Pour toute visualisation en général, quels sont les risques encourus en cas de mauvais choix des couleurs ? **[1 pt]**
3. Rappeler la définition de « l'efficacité » pour des visualisations. **[1 pt]**
4. Les chercheurs ont démontré expérimentalement que la méthode 1 est plus efficace que la méthode 2 pour des tâches d'étude de détail sur les séries, comme par exemple comparer la valeur des séries à une date donnée. La méthode 2 semble plus efficace pour des tâches d'études globales sur les séries, comme par exemple trouver la série qui fluctue la plus. Donner un autre exemple d'étude de détail qui est facilitée avec la méthode 1 et au contraire difficile avec la méthode 2. **[1 pt]**
5. Donner un autre exemple d'étude globale qui est facilitée avec la méthode 2 et au contraire difficile avec la méthode 1. **[1 pt]**

D'après : W. Javed, B. McDonnel, N. Elmqvist. *Graphical Perception of Multiple Time Series. IEEE Trans. Visualization and Computer Graphics*, vol. 16(6), pp. 927-934, 2010.

Correction :

1. Il s'agit d'un cas particulier des « petits multiples » de Tufte que ce dernier appelle des « sparklines » (cf exercice dirigé consacré au sujet).
2. Le premier risque est une mauvais perception par l'utilisateur, du fait d'une déficience possible des cônes de la rétine oculaire (daltonisme, par exemple). Le cas concerne 10% environ des hommes, c'est donc loin d'être négligeable. L'autre risque concerne le dispositif de visualisation. Si celui de l'utilisateur est différent de celui

du concepteur de la visualisation, il y a un risque de non reproduction de la couleur, du fait des gamuts différents des appareils.

3. Celle-ci est définie par Bertin (cf partie du cours sur la perception) : « Si pour obtenir une réponse correcte et complète à une question donnée, et toutes choses égales, une construction requiert un temps d'observation plus court qu'une autre construction, on dira qu'elle est plus efficace pour cette question ».
4. Il est bien plus facile de trouver quand les séries ont une valeur commune avec la méthode 1 car cela correspond à une intersection des courbes, très simple à visualiser.
5. Déterminer quelles sont les courbes qui sont majoritairement à forte (ou faible) valeur est clairement facilité par la méthode 2.