le cnam

Sujet UE RCP216

Ingénierie de la fouille et de la visualisation des données massives

Année universitaire 2020–2021

Examen 1ère session : 26 janvier 2021

Responsable : Michel CRUCIANU

Durée : 2h00

Aucune communication n'est autorisée durant l'examen à l'exception des échanges avec les serveurs du Cnam et avec les enseignants de cette unité d'enseignement.

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 (10 points)

[4 pts] Question 1 : Passage à l'échelle par distribution.

1. Pourquoi la distribution est la solution préférée aujourd'hui pour traiter de très grands volumes de données ? (**1 pt**)
2. Quels mécanismes sont mis en œuvre pour que les calculs distribués soient résilients aux pannes de nœuds de calcul ? (**2 pts**)
3. Pourquoi faut-il minimiser les échanges avec le stockage de masse (disque) entre les étapes (ou les itérations) successives d'un calcul ? (**1 pt**)

Correction :

1. La distribution permet (avec certaines restrictions concernant les réseaux de communication et les algorithmes mis en œuvre) d'augmenter facilement les capacités de traitement (presque linéairement avec le nombre de nœuds de calcul), sans « goulot d'étranglement » particulier.
2. Un mécanisme de réplication permet de stocker chaque donnée sur plusieurs ordinateurs afin de réduire le risque de perte malgré des pannes individuelles. L'interruption des calculs en cours dans le système distribué, avec obligation de reprise depuis le début, est évitée en partitionnant les calculs en « grains » traçables et suffisamment fins pour qu'en cas de panne les grains abandonnés puissent être efficacement réaffectés aux ordinateurs encore actifs.
3. Les échanges avec le stockage de masse ralentissent très significativement le déroulement d'un algorithme, voir la Fig. 3 du cours 1.

[3 pts] Question 2 : Systèmes de recommandation (SR).

1. Les SR par similarité de contenu utilisent-ils des informations complémentaires par rapport aux SR par filtrage collaboratif ou les mêmes informations ? Indiquez de quelles informations il s'agit. (**1 pt**)
2. Pour le filtrage collaboratif basé sur la factorisation matricielle, quel est l'intéret de la modélisation d'un biais par utilisateur ? A votre avis, qu'est-ce qui conditionne une estimation fiable de ces biais ? (**2 pts**)

Correction :

1. A la fois des informations de même nature, celles issues de la matrice d'utilités, et des informations complémentaires, concernant des caractéristiques des articles et éventuellement des utilisateurs.

2. Les niveaux d'exigence des différents utilisateurs ne sont pas les mêmes, les notes données par différents utilisateurs à un même article ne peuvent donc pas être comparées directement. Pour une estimation fiable des biais par utilisateur il est nécessaire d'avoir suffisamment de données dans la matrice d'utilités. Par ex., on ne peut pas bien estimer le biais d'un utilisateur s'il a noté un seul article et est le seul à avoir noté cet article.

[3 pts] Question 3 : Flux de données et modélisation incrémentale.

1. Comment peut-on définir (brièvement) la modélisation incrémentale à partir de données ? (**1 pt**)
2. Indiquez deux intérêts majeurs de la modélisation incrémentale pour le traitement de flux de données. (**2 pts**)

Correction :

1. La modélisation incrémentale (à partir de données) consiste à actualiser un modèle (descriptif ou décisionnel) au fur et à mesure que de nouvelles données sont disponibles, sans avoir à le ré-estimer complètement avec la totalité des données.
2. Un premier intérêt majeur de la modélisation incrémentale pour le traitement de flux de données est la possibilité de mettre à jour un modèle progressivement, en tenant compte à chaque fois des dernières données arrivées, de faible volume donc, et par conséquence avec un faible coût de calcul et sans avoir à stocker toutes les données depuis le début du flux (pour pouvoir ré-estimer complètement le modèle). Un second intérêt majeur est la facilité à privilégier les données récentes par rapport aux plus anciennes, donc à tenir compte du caractère souvent non stationnaire de la distribution des données.

2 Fouille de graphes et réseaux sociaux (4 points)

[3pts] Question 4 : On cherche à étudier la diffusion de l'information au sein d'une entreprise. Pour ce projet, vous collectez tous les courriels échangés dans l'entreprise durant la dernière année sur la base du volontariat. Vous modélisez ensuite le graphe orienté des échanges : chaque courriel est une arête reliant la personne qui l'a expédié à son destinataire (les nœuds sont donc des personnes). On considère un graphe pondéré : le poids d'une arête correspond au nombre de courriels ayant transité entre ses deux extrémités.

1. Donnez deux exemples de biais liés à la façon dont les données ont été collectées pouvant influer sur votre analyse de ce réseau social.
2. Quelle signification donner aux communautés obtenues sur ce réseau social (par exemple, à l'aide de l'algorithme de Louvain) ?

3. Proposez une façon de visualiser ce graphe de sorte à mettre en avant les communautés et les échanges les plus fréquents. En particulier, suggérez un algorithme de spatialisation qui vous semble adapté et une façon adéquate de représenter les nœuds et les arêtes de ce réseau.

Correction :

1. Le principal biais de cette analyse est qu'elle ne prend pas en compte les échanges d'information hors courriel (réunions, discussions privées, etc.). D'autres biais peuvent survenir, par exemple :
   — Un courriel reçu n'est pas forcément lu.
   — Certains courriels sont générés automatiquement (listes de diffusion, répondeur automatique…).
   — Les volontaires ne sont pas forcément représentatifs de l'organisation dans son entièreté.
   — Certaines communications peuvent être exclues d'office car il ne serait pas légal de les collecter (courriels syndicaux ou adressés aux déléguées du personnel).
2. Les communautés obtenues représentent des groupes de personnes qui échangent plus souvent des courriels entre elles qu'avec le reste du réseau. On s'attend par exemple à ce qu'un département ou une équipe de l'entreprise forme une communauté.
3. Une spatialisation circulaire peut être adaptée pour rendre particulièrement visible les communautés. Les algorithmes de spatialisation type ForceAtlas peuvent permettre de rendre visible les *hubs* (des personnes centrales). Pour les nœuds, on peut les colorer en fonction de la communauté d'appartenance et grossir le nœud en fonction de sa centralité ou de son degré (montre l'importance d'une personne dans le graphe des courriels). Pour les arêtes, on peut les colorier ou les épaissir en fonction de leur poids (le nombre de courriels échangés).

[1pt] Question 5 : Considérons le graphe suivant (Figure 1). Décrire l'ordre de visite des nœuds en suivant un parcours **en largeur** à partir de A.

Correction :

$$A \to B \to D \to C \to G \to E \to F$$

3 Visualisation et interaction (6 points)

[4 pts] Question 6 : La figure ci-dessous représente un graphe sous forme de boites pour les sommets et de lignes pour les arcs. On sait que cette représentation a un mauvais ratio data/encre.

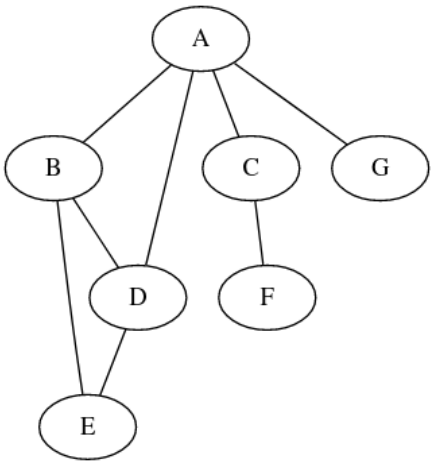
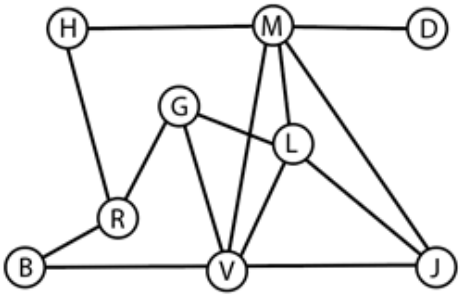


FIGURE 1 – Exemple de graphe



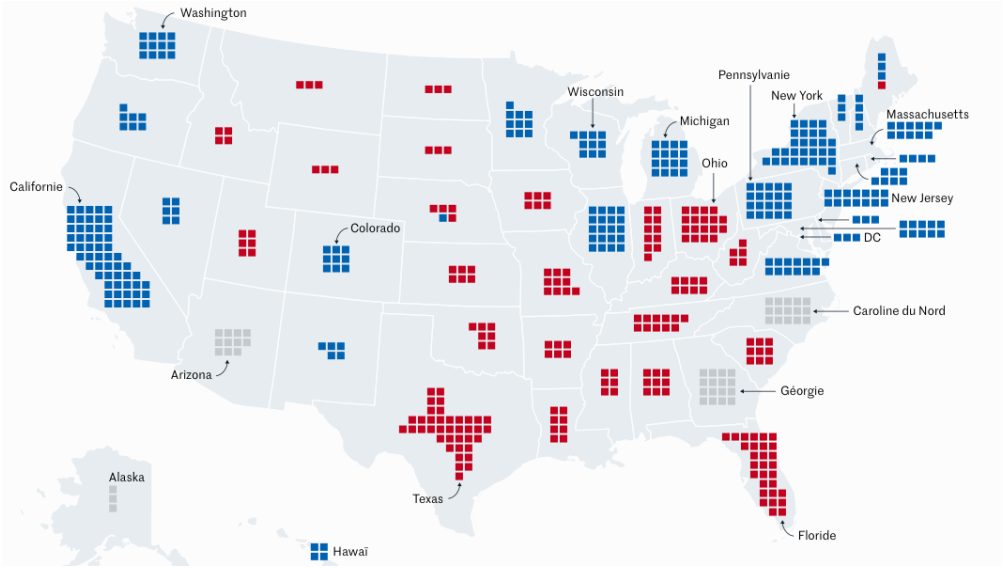
1. Définir le ratio data/encre d'une représentation. Pourquoi est-il important de chercher à maximiser le ratio data/encre d'une représentation ? (2 points)
2. Citer d'autres représentations de graphes qui ont un meilleur ratio data/encre que la représentation boites-lignes. (2 points)

Correction :

1. Le ratio data/encre est un concept développé par E. Tufte. Pour une visualisation sur écran, il s'agit de compter le nombre de pixels directements utilisés pour les données et ceux relevant plutôt de la décoration (comme les axes gradués des courbes). Maximiser le ratio permet de densifier les visualisations (c.a.d. afficher plus de données pour une surface de visualisation donnée).
2. Le *treemap*, la matrice de précédence et, dans une moindre mesure, les méthodes mixtes comme NodeTrix. En éliminant le dessin des arcs, ces techniques augmentent le ratio data/encre.

[2 pts] Question 7 : Les récentes élections américaines ont donné lieu à de très nombreuses productions de cartes en ligne. La carte ci-dessous est extraite du journal Le

Monde (8 nov. 2020). Elle indique, pour chaque état, le nombre de grands électeurs de cet état et leur parti politique (bleu pour démocates, rouge pour républicains, gris pour les résultats en attente). On voit que les états ont un nombre très variable de grands électeurs : des dizaines dans l'état de Californie, trois dans les états du centre nord.



Quelle transformation de la carte permettrait de mieux tenir compte de la grande variation du nombre de grands électeurs entre les états ?

Correction : Une distorsion (anamorphose) qui affecterait à chaque état une surface proportionnelle au nombre de grands électeurs. Voir les exemples du laboratoire Choros à l'EPFL cités en cours.