le cnam

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

Année universitaire 2016–2017

Examen 1ère session : 27 juin 2017

Responsable : Michel CRUCIANU

Durée : 3h00

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.

Sujet de 8 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.

Fouille de données (13 points)

[2 pts] Question 1 : Quels avantages présente l'utilisation d'un algorithme d'initialisation de type *K-means||* par rapport à une initialisation aléatoire des centres avant application d'une classification par *K-means* ?

Correction :

— La solution obtenue est *meilleure* : pour un même nombre de groupes, le minimum atteint correspond en général à une valeur plus faible de la somme des inerties intra-classe des groupes.

— Avec une initialisation de type *K-means||*, la convergence de *K-means* nécessite moins d'itérations et est donc *plus rapide*.

[2 pts] Question 2 : Pourquoi la factorisation régularisée est préférable à la décomposition en valeurs singulières pour le filtrage collaboratif ?

Correction :

La matrice utilisateurs-articles, à factoriser, est très *incomplète*, la décomposition en valeurs singulières (SVD) ne peut donc pas être appliquée. Donner une valeur particulière (par ex. 0) aux données manquantes pour pouvoir appliquer la SVD n'a pas de justification, l'objectif est justement d'obtenir un modèle pour prédire les valeurs manquantes.

[2 pts] Question 3 : Expliquez pourquoi avec un filtre de Bloom il n'y a pas de faux négatifs (données qui auraient dû être filtrées mais ne le sont pas) mais, en revanche, nous pouvons avoir de faux positifs (données qui n'auraient pas dû être filtrées mais qui l'ont été).

Correction :

Les vrais positifs (données qui doivent être filtrées) font partie d'une liste connue à l'avance et, *par construction*, un filtre de Bloom les détecte comme des positifs (il n'y a donc pas de positifs non détectés, c'est à dire de faux négatifs). En revanche, rien n'est fait pour *ne pas* filtrer les vrais négatifs (données qui ne doivent pas être filtrées). Par ailleurs, les vrais négatifs ne sont pas nécessairement tous connus à l'avance.

[3 pts] Question 4 : Quel intérêt de principe présente l'emploi de représentations Word2Vec

par rapport à l'utilisation de représentations de type LSA dans la fouille de textes ? Quelles sont les contraintes associées à l'emploi de représentations Word2Vec ?

Correction :
LSA est basée sur une représentation d'un texte comme un *ensemble* de mots et ignore ainsi toute information de contexte pour les mots, information exploitée par Word2Vec. Parmi les contraintes de Word2Vec nous pouvons mentionner la nécessité de disposer d'un volume important de textes (de même nature et de même vocabulaire que les textes à analyser) pour construire des représentations Word2Vec de bonne qualité, ainsi que la nécessité de se limiter à l'analyse de textes (très) courts car chaque texte est représenté par le centre de gravité des représentations Word2Vec de ses mots.

[1 pt] Question 5 : Sur le graphe présenté figure 1, quel est le nœud qui présente la plus forte centralité d'intermédiarité ? Justifiez votre réponse, sans présenter de calcul.

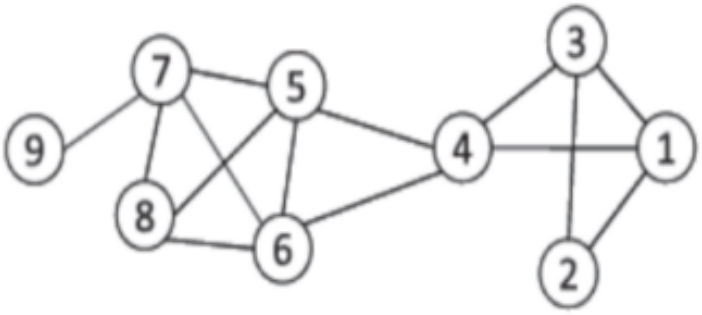

FIGURE 1 – Graphe

Correction :
Le nœud 4 sera sur le plus grand nombre de plus courts chemins, c'est le nœud de plus forte centralité.

[1 pt] Question 6 : Quelles propriétés attend-on des graphes de terrain ? En quoi diffèrent-ils significativement de graphes aléatoires ?

Correction :
Coefficient de clustering, effet petits mondes, densité, diamètre, plus courts chemins.

[2 pts] Question 7 : Quels sont les avantages et inconvénients de l'algorithme détection de communautés de Louvain ? Citez au moins deux avantages et deux inconvénients.

Correction :

— Avantages : vitesse, optimisation de la modularité (qualité), stockage efficace, passage à l'échelle ;

— Inconvénients : limite de résolution, non déterminisme.

Visualisation et interaction (7 points)

[2 pts] Question 8 : Mettre des couleurs dans un graphique pour distinguer des catégories de données est efficace mais risqué.

1. (**1 point**) Pourquoi est-ce efficace ? En particulier, quel est le phénomène cognitif à l'œuvre ?
2. (**1 point**) Donner 2 raisons qui risquent de rendre le coloriage inefficace pour l'utilisateur.

Correction :

1. C'est efficace grâce à notre capacité de vision pré-attentive. Des formes de couleur différentes sont distinguées par l'humain sans effort cognitif, donc très vite.
2. La première est que tous les humains ne reconnaissent pas les couleurs de la même façon, à cause d'une répartition inégale des cellules de type cône qui permettent la distinction des couleurs (par exemple, daltonisme du fait de l'absence de cônes sensible au vert. La deuxième est que les dispositifs de restitution (écrans, projecteurs vidéo, imprimantes, etc.) ont une capacité limitée pour reproduire une couleur, définie par leur gamut. Ce gamut défini un triangle dans le diagramme CIE avec les trois couleurs de base utilisées par le dispositif (RVB, par exemple). Seules les couleurs à l'intersection des triangles subsiste d'un dispositif à l'autre. On peut toutefois construire des visualisations qui résistent à ces deux problèmes et il existe plusieurs sites web qui informent à ce sujet (cf cours 2).

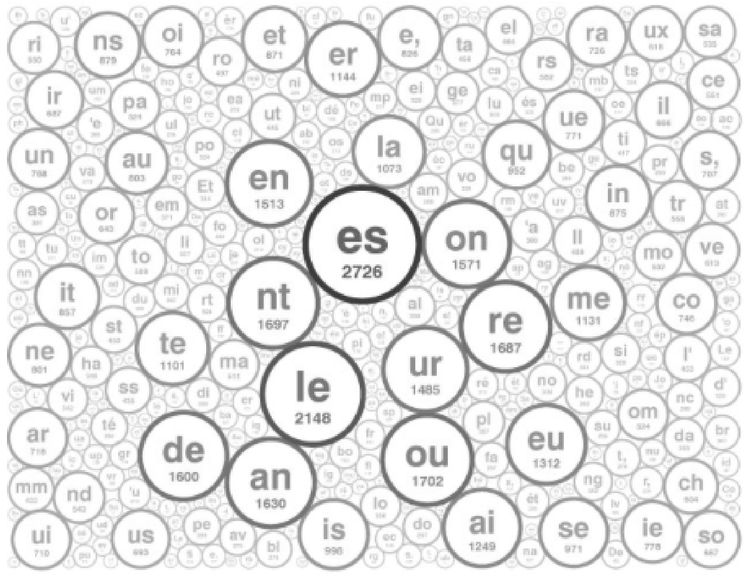

FIGURE 2 – Un exemple de *bubble chart* occurences des digrammes dans un texte français. Extrait de https ://www.m-i-b.com.ar/letters/en/

[4 pts] Question 9 : On appelle *bubble chart* les graphiques du type de celui reproduit ci-dessous. Ici, les données quantitatives sont visualisées sous la forme de cercles dont la surface est proportionnelle à la valeur des données. Le placement des cercles est sans importance, il s'agit surtout d'obtenir le graphique le plus compact possible.

1. (**1 point**) Quelle est la variable rétinienne de la liste de Bertin utilisée de ces graphiques ?
2. (**1 point**) Selon Mackinkay, quelle est la variable rétinienne la plus efficace qui pourrait être utilisée dans ce contexte ? Quel autre genre de graphique obtient-on dans ce cas ?
3. (**2 points**) En s'inspirant d'algorithmes vus en travaux pratiques, décrire les grandes lignes d'un algorithme de dessin de *bubble charts* qui réalise le placement automatique des cercles sur une surface de dimensions fixées à l'avance. On ne demande pas un programme fonctionnel.

Correction **:**

1. Bertin regroupe sous le terme de "taille" toutes les techniques qui consistent à prendre une forme dont on fait varier les dimensionsLes expériencxes de Clevelenad et McGill distinguent plus précisément les variations de taille : variation de longueur, variation de surface etc.
2. Dans le tableau de Mackinlay, on voit que pour des données de type quantitatif, l'attribut de longueur est plus performant que celui de surface. On obtient dans ce cas un simple graphique en barres, moins spectaculaire sans doute. (remarque : le *bubble graph* favorise sans doute mieux la découverte accidentelle d'information, en revanche).
3. On a vu en TP un problème très similaire avec le dessin de nuages de mots (tag-cloud). L'algorithme consistait en :

   1) Classer les formes par dimensions, dans une liste. On va traiter cette liste de la plus grosse forme à la plus petite. Ici les formes seront des cercles, de rayon plus ou moins grand. On attribue une première position à chacune des formes dans le dessin au hasard (répartition uniforme avec la fonction random).

   2) dans une boucle, faire un parcourt de toute la liste classée : - examiner si la forme en cours intersecte avec tous les suivants - si non, passer à la forme suivante dans la liste - si oui, donner une nouvelle position au hasard à la forme qui gène la forme en cours d'examen

   3) On sort de la la boucle quand il n'y a plus d'intersections ou quand le compteur de tours a dépassé sa limite.

   Cet algorithme est simple à mettre en oeuvre mais bien sûr n'est qu'une heuristique pour ce problème de placement fortement combinatoire.

[1 pt] Question 10 : La carte de France reproduite ci-dessous a été calculée par le laboratoire de géographie de l'Ecole Polytechnique Fédérale de Lausanne. Elle n'utilise pas de projection géographique standard. Ici, chaque zone administrative est dessinée avec une surface proportionnelle à sa population. Le poids de l'Ile de France est de ce point de vue frappant.

1. (**0,5 points**) Donner un exemple de données pour lequel ce type de représentation permet une meilleure interprétation que la carte traditionnelle.
2. (**0,5 points**) Parmi les techniques de représentations vues en cours, à laquelle se rapporte ce genre de carte ?

Correction :

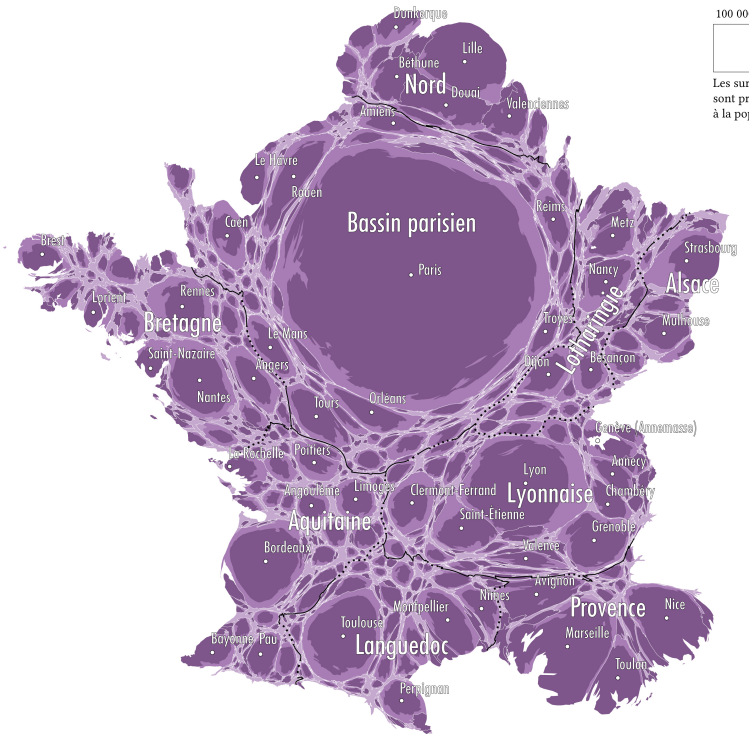

FIGURE 3 – (extrait de http ://choros.epfl.ch/)

1. Les élections ! C'était d'actualité cette année. Les journaux montrent très souvent des cartes de France avec une couleur type par département selon le vote majoritaire. Tenir compte de la population du département donne une bien meilleure image des forces en présence, car les départements ruraux très peu peuplés sont sinon trop visibles.
2. Les technique dites de distorsion (cf cours 4 interaction), en particulier les anamorphoses.
