

Recherche de communautés

Cours Master 2, février 2010

Michel Habib

habib@liafa.jussieu.fr

<http://www.liafa.jussieu.fr/~habib>

février 2010

Plan

Classification, Clustering

Méthodes de partitionnement / recouvrement

- ▶ Il existe de nombreuses données sur des graphes d'interaction disponibles
- ▶ Comment trouver des communautés dans ces graphes ?
- ▶ **Explicites**
Exemple : les anciens élèves du Lycée Jacques Decour de 1990 à 2000, de la légion étrangère de 1945 à 1960, de l'UFR informatique Université Paris Diderot, du master SRI .
Les membres ont conscience d'appartenir à cette communauté.
Francs-maçons et députés
- ▶ **Implicites** les membres n'en ont pas forcément pris conscience.
Ces communautés sont obtenues par calcul à partir de données, notion plus intéressante algorithmiquement.
Exemple : Députés ayant voté de la même façon

- ▶ Recherche de communautés par émergence.
- ▶ Exemple : Sous-disciplines scientifiques par analyse des citations
- ▶ A des fins de Marketing

Définitions

- ▶ Mettre ensemble ceux qui se ressemblent
- ▶ Détecter, extraire les ensembles de sommets qui jouent des rôles identiques (ou presque identiques) dans un réseau social (i.e. les communautés explicites ou implicites).
- ▶ Mettre ensemble ceux qui ont les même interactions

Les modules

- ▶ $M \subseteq X$ est un module ssi
 $\forall x, y \in M, N(x) - M = N(y) - M$
- ▶ La décomposition introduit une arborescence de décomposition
- ▶ Hélas la majorité des graphes sont indécomposables
- ▶ La définition se généralise difficilement

Les rôles

- ▶ La notion de rôle issue des théories des réseaux sociaux années 1970.
- ▶ Deux sommets jouent le même rôle ss'ils ont les mêmes couleurs dans leur voisinage
- ▶ Généralisation de la notion de module
- ▶ Très bonne idée, mais hélas le Problème est NP-complet (Fiala 1995)

Deuxième formalisation

- ▶ Etant donné un graphe, il s'agit de trouver une partition des sommets en classes qui vérifient :
- ▶ Beaucoup de relations intra-classe
- ▶ Peu de relations interclasses
- ▶ Idée générale : les liens sont plus forts entre les éléments de la communauté que les liens intercommunautaires (les liens peuvent être valués)

Clustering

Nom : Max-Cut

Données : un graphe $G = (X, E)$ une valuation $\omega : E \rightarrow N$, un entier positif k

Question : Existe-t-il une partition de X en X', X'' telle que la somme des valuation des arêtes entre X' et X'' soit supérieure à k ?

Théorème :

Max-Cut est NP-complet.

Approximabilité

Théorème :

Max-Cut est approximable avec un facteur 1,1383 par Goemans et Williamson (1995).

Pas approximable avec un facteur de 1,0624 (Hastad, 1997)

Théorème :

Max-Cut est APX-complet Papadimitriou et Yannakakis (1991).

Cas d'un graphe orienté

Les résultats sont analogues (les coefficients changent un peu).

- ▶ C'est un problème que l'on rencontre dans de très nombreux domaines (image, recherche opérationnelle, ...)
- ▶ On l'appelle aussi classification ou clustering
- ▶ Sujet très étudié

Méthodes bottom up

- ▶ Les méthodes par croissance de classes sont très utilisées en analyse d'image et en statistique pour analyser des nuages de points.
- ▶ Surtout utilisé en imagerie
- ▶ fusion de zones de couleur voisines
- ▶ En général la solution dépend de l'ordre dans lequel on a analysé les pixels de l'image

Méthodes top-down

S'il existe un critère de séparation pertinent

Méthodes exotiques

- ▶ Modèles particuliers avec attraction : Cosmoweb (Bennouas, Bouklit, de Montgolfier)
- ▶ Recuit simulé, méthodes Tabou, méthodes génétiques
- ▶ Analogie électrique
- ▶ Méthodes à partir de marches aléatoires (l'idée est qu'au début la marche aléatoire reste dans la communauté)
- ▶ Extraction des liens les plus utilisés dans les plus courts chemins
- ...

- ▶ Le problème est approximable polynomialement
- ▶ L'heuristique choisie doit dépendre du domaine d'application

Nuées dynamiques, K-moyennes

1. Principe des nuées dynamiques

On dispose de l'ensemble des couleurs de l'image à quantifier. Cet ensemble constitue un nuage de couleurs dans l'espace colorimétrique utilisé. On va alors réaliser une agrégation en amas de couleurs. Cette classification se base sur la minimisation d'une quantité représentant la notion de dispersion des classes de couleur (= des points dans l'espace des couleurs).

2. Algorithme des nuées dynamiques

Les nuées dynamiques sont en fait une généralisation de l'algorithme de clustering bien connu des k-moyennes. La généralisation des k-moyennes se situe dans l'utilisation d'une mesure plus générale de dissemblance d'un point à sa classe.

L'algorithme

1. Initialisation : Choisir les centres initiaux des classes.
2. Affectation : Calcul de la classe de chaque point du nuage de couleur.
3. Mise à jour des centres (ou attributs des classes) : Calcul des nouveaux centres de chaque classe (couleur barycentre de toutes les couleurs de la classe)
4. Terminaison basée sur : le nombre d'itérations ou la stabilité des positions des centres

- ▶ Optimisation locale d'une solution par échanges
- ▶ Benchmarks faciles à construire

Applications aux algorithmes de recommandation

- ▶ Recherche des bicliques (ou quasi-bicliques) maximales pour l'inclusion.
Biclique= sous-graphe biparti complet
- ▶ stars versus fans
- ▶ clients versus livres (Amazon)
- ▶ clients versus produits
- ▶ Réseaux sociaux
- ▶ Il s'agit de la recherche de sous-graphe bipartis complets maximaux que l'on identifie à des communautés. Problème algorithmiquement difficile.

- ▶ Le problème de la recherche d'une biclique de cardinal maximum dans un graphe biparti est NP-complet.
- ▶ L'énumération des bicliques maximales dans un graphe biparti est un problème $\#$ P-complet.
- ▶ Il n'existe donc que des heuristiques polynomiales ou des algorithmes exacts exponentiels.

Problème de recherche

Peut-on utiliser dans l'algorithme la structure du graphe que l'on considère ?

Par exemple : graphe petit monde, graphe de terrains

Pour tester les heuristiques, il suffit de produire des graphes construits avec une décomposition et de tester si l'heuristique la trouve.

Etude des voisinages de sommet, recherche de régularités
Typologie des sous-graphes, .