Propagation de label

Loïc Maurin & Samuel Rincé

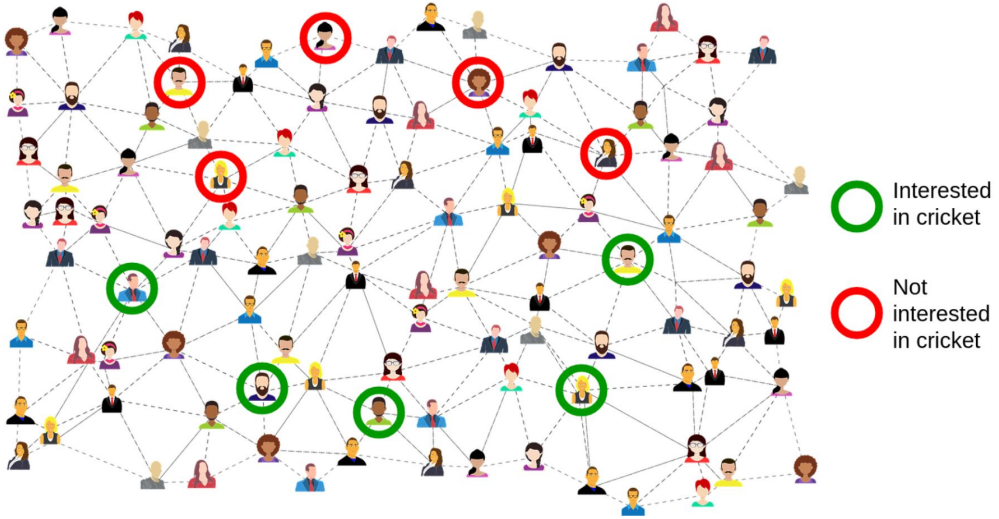
GitHub: https://github.com/samuelrince/labelpropagation

Sommaire

1. Problème de propagation de label
2. Algorithme
3. Implémentation avec Pregel
4. Datasets
5. Expériences
    a. Influence de la quantité de labels masqués
    b. Influence du nombre de noeuds sur le temps de calcul
6. Conclusion

Problème de propagation de label

- Classification de caractéristiques (appartenance à un parti, etc.) par relations sur un graphe.
- Apprentissage semi-supervisé
- Méthodes
  - Random Walk
  - Zhu et al.
  - Classification Bayésienne



Algorithme

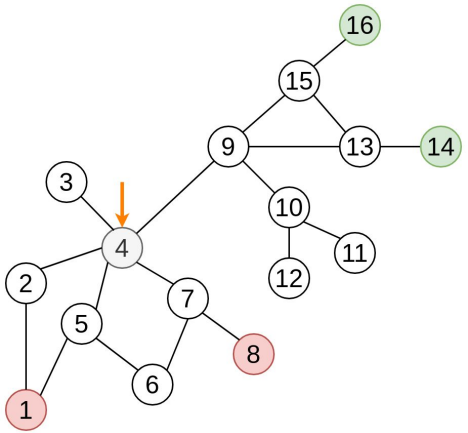
- Publication [Zhu et al. 2002]
- Calcul des labels avec par la méthode itérative
- Convergence contrôlée par MAX_ITER

compute $D_{ii} = \sum_j A_{ij}$
compute $P = D^{-1}A$
$Y_0 = (Y, 0)$
$t = 0$
while $t <= MAX\_ITER$ **do**
    $Y^{t+1} \leftarrow PY^t$
    $Y^{t+1} \leftarrow Y^t$
end
$Y\_final = Y^t$

Implémentation avec Pregel

- Description du graphe :
  - VertexAttribute
    (degree, isInitialLabel, LabelsArray)
- Vertex Program :
  - Si label initial : Y reste inchangé
  - Sinon : Y = Y / degree
- Send Message :
  - Labels
- Merge Message :
  - sum(Labels[neighbors])

- Converge en un nombre d'itérations de l'ordre de grandeur du diamètre du graphe.



Datasets

- Dataset de test : `soc-karate`
  - Lien entre des clubs de karaté
  - 34 Noeuds — 78 Arêtes

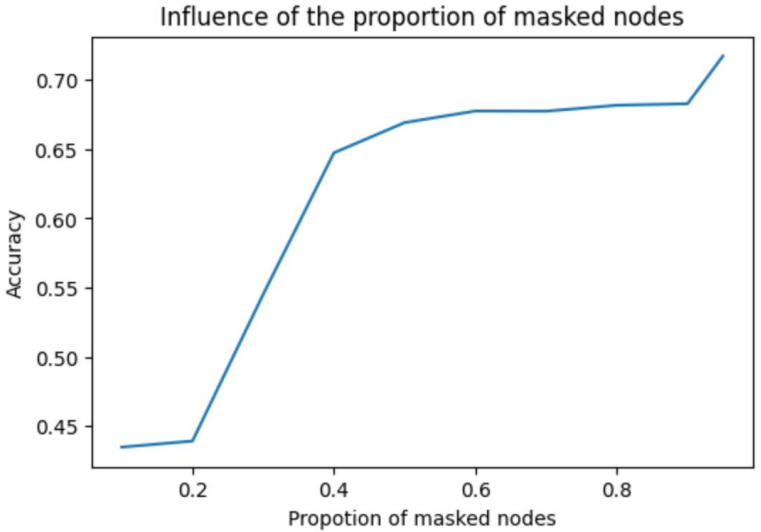
- Dataset d'expérimentation : `email-Eu-core`
  - 1k Noeuds — 25k Arêtes
  - Lien entre les laboratoires de recherche ayant communiqué par emails en Europe

- Dataset d'expérimentation : `com-DBLP`
  - 320k Noeuds — 1M Arêtes
  - Dataset de lien entre des papiers de recherche scientifique
  - Par manque de RAM nous avons dû l'abandonner

Sommaire

Influence de la quantité de labels masqués

- Observé les différences de labellisation en fonction du nombre de noeuds labellisé initialement
- Masquage aléatoire à *seed* fixé
- Mesure de la performance avec l'*accuracy* de classification
- Calcul à MAX_ITER constant choisi très grand



Influence du nombre de noeuds sur le temps de calcul

- Sur soc-karate :
  - 2 s environ
- Pour un graphe initial de 1000 noeuds :
  - 70 s environ
- Améliorations possibles :
  - Essais sur le nombre d'arêtes
  - Prise en compte du rayon et du diamètre du graphe

Nbr noeuds	Temps de calcul
1000	77 s
750	71 s
500	76 s
250	68 s
100	68 s

Conclusion

- Peu d'expériences ont pu être menées.
- Les variations des hyperparamètre n'impliquent pas de grande variation des résultats
    - Application des algorithmes sur des graphes à trop petite échelle
    - Manque de capacité de calcul pour passer à l'échelle
    - Optimization de notre algorithme possible
- Difficultés d'implémentations