

Application des filtres collaboratifs pour site de rencontre

Alexandre Spaeth

18 nov. 2009

Introduction

Contexte

Intérêt du projet

Méthode

Méthode initiale

Méthode améliorée

Expériences

Quantité des recommandations

Qualité des recommandations

Résultats

Jeux de données utilisés

Quantité des recommandations

Qualité des recommandations

Conclusion

Avenues de recherche

Contexte de la recherche

- ▶ Les filtres collaboratifs sont très utilisés dans les domaines commerciaux
- ▶ Les applications aux sites de rencontres en ligne ne sont pas très répandus
- ▶ Les exemples connus : <http://colfi.wz.cz/> et <http://www.okcupid.com> mais les votes sont explicites

Objectif du projet

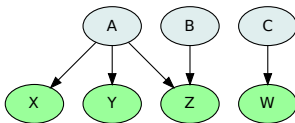
- ▶ Appliquer les théories des filtres collaboratifs aux sites de rencontre en ligne
- ▶ Se contenter autant que possibles de données existantes de navigation (votes implicites)
 - ▶ Pas de période de collecte de données
- ▶ Explorer une nouvelle approche pour les profils avec peu de données (problème du *cold-start*)
 - ▶ Utiliser notamment les données démographiques (âge, education, ethnicité, etc.)

Données utilisées

- ▶ On base notre recommandation sur les données implicites collectées par le site :
 - ▶ Fiches-profils vues
 - ▶ Liste des favoris
 - ▶ Clins d'œil
- ▶ On pondère chaque élément par une valeur différente (par exemple un clin d'œil = 4 fiches vues)
- ▶ On crée un graphe des relations entre les utilisateurs que l'on utilise ensuite.

Méthode item-item

- ▶ Le nombre de liens dans le graphe est faible, ce qui permet d'avoir des matrices d'adjacence très creuses (moins de .1% de liens)
- ▶ Considérons le graphe de relations suivant :



- ▶ On peut recommander le profil X à B mais rien à A ou à C.

Justification

La méthode explicitée précédemment correspond à repérer les chemins $B \leftarrow Z \rightarrow A \rightarrow Y$. En considérant la matrice d'ajacence $M_{i,j}$, où un lien $i \rightarrow j$ correspond à un vote, alors :

$M \cdot M^T$ donne le nombre de votes communs pondéré par la valeur des votes.

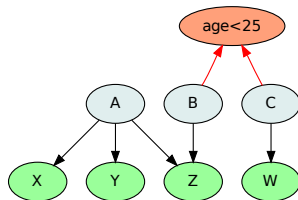
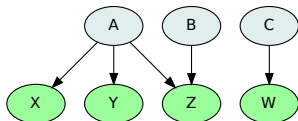
$M^T \cdot (M \cdot M^T)$ donne le nombre de chemins $B \leftarrow Z \rightarrow A \rightarrow Y$.

En pratique, on utilise le *cosinus* plutôt que le produit scalaire. La formule utilisée finalement est donc :

$$M^T \cdot \frac{M \cdot M^T}{\|M \cdot M^T\|}$$

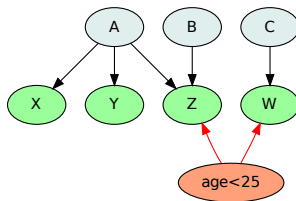
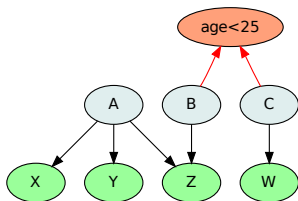
Problème des profils nouveaux

- ▶ Dans le cas où l'historique du profil n'est pas assez fourni, on ne sera pas capable de faire de recommandation.
- ▶ Il faut donc créer de nouveaux liens en utilisant les données que l'on a : les données démographiques.



Comparaison à l'autre solution possible

- ▶ On pourrait ajouter les profils des cibles
- ▶ Mais alors il faudrait minimalement que chacun ait un vote, ce qui n'est pas toujours vrai
- ▶ En pratique, le calcul est plus facile à gérer (temps et mémoire) dans le sens que l'on a choisi
- ▶ Le problème consiste alors à choisir les traits à conserver



Combien de recommandations peut-t-on faire ?

- ▶ On cherche à déterminer le nombre de personnes pour lesquelles il est possible de faire une recommandation en se basant sur la méthode item-item classique.
- ▶ À noter que pour un trop grand nombre de données, on est obligé de diviser les données en bloc aléatoires.
- ▶ Afin de ne pas perdre de recommandations possibles, on va garder en mémoire les personnes sans recommandations et on va les inclure dans d'autres blocs.

Quelle est la qualité de ces recommandations ?

- ▶ Une solution coûteuse est de faire réellement les recommandations et de constater le taux de conversion.
 - ▶ C'est trop long !
 - ▶ C'est risqué si les recommandations sont mauvaises.
- ▶ Une autre solution est de faire une validation croisée
 - ▶ Ce n'est pas très efficace car le nombre de votes manquants est beaucoup trop important
- ▶ Enfin, on peut aussi utiliser les historiques de recommandations qui ont été faits jusqu'ici.
 - ▶ On fait des recommandations.
 - ▶ On vérifie dans l'historique si c'est une recommandation déjà faite
 - ▶ Si c'est le cas, on regarde si elle était bonne ou non
 - ▶ On fait la moyenne à la fin pour calculer le taux de prédictions correctes

Évaluation des 2 méthodes

- ▶ Il faut évaluer les 2 méthodes pour savoir le nombre de personnes à qui on peut faire une recommandation et pour connaître la qualité des prédictions.
- ▶ En pratique, la 2^e méthode n'est utilisée que pour les personnes ne disposant pas de recommandations avec la première (calcul trop lourd sinon).

Jeux de données

- ▶ Le jeu de données utilisé correspond aux historiques de navigation et de discussion de 141'778 personnes.
- ▶ Le nombre de cibles potentielles est alors de 658'225 personnes.
- ▶ Cela correspond à 3'733'709 évènements (ou votes implicites)
- ▶ En moyenne, chaque personne a voté pour 26 profils.
- ▶ 35'014 personnes n'ont aucun vote.

Nombre de recommandations possibles

- ▶ Avec la première méthode, on ne peut effectuer des recommandations qu'aux personnes ayant un historique.
- ▶ On visait donc à effectuer des recommandations pour $141'778 - 35'014 = 106'764$ personnes.
- ▶ 106'173 personnes ont pu obtenir une recommandation (soit plus de 99% des cibles).
- ▶ Ainsi, on peut faire des recommandations pour des personnes avec un seul vote.
- ▶ Cela prend environ 75 minutes pour effectuer ces recommandations (1400 fiches par min.)

Application de la deuxième méthode

- ▶ Dans le cas où on ne trouve pas de recommandations, on ajoute les liens de profil.
- ▶ On arrive alors à trouver une recommandation pour pratiquement 100% des personnes
- ▶ Mais c'est beaucoup plus lourd car il faut 67 min pour les 36'000 personnes restantes (500 fiches par min.)

Et la qualité ?

- ▶ Pas de résultats précis à donner pour l'instant par manque de données !!
- ▶ Les premiers tests semblaient très encourageants cependant.

Conclusions

- ▶ On peut faire des recommandations pour des personnes avec peu d'historique
- ▶ Il est possible d'inclure les données démographiques dans le graphe des relations
- ▶ Les premiers résultats sont encourageants

Avenues de recherche

- ▶ Quelle valeur donner à chaque vote implicite ?
- ▶ Quels sont les traits démographiques à conserver ?
- ▶ Y a-t-il une méthode efficace d'évaluation des recommandations ?
- ▶ Peut-t-on utiliser les données psych-sociales ?
- ▶ Est-il possible d'intégrer les données de profil des liens visités sans que cela soit trop lourd à calculer ?