Filtrage collaboratif (Systèmes de recommandations)

Brice Mayag

M1 SIREN

Plan

Filtrage collaboratif: Définition

But:

- Permet d'automatiser la sélection et la recommandation d'articles (filtrage);
- Sert à prédire les préférences d'un utilisateur à partir des préférences d'autres utilisateurs (collaboratif).

Filtrage collaboratif: Applications et enjeux

Les systèmes de Filtrage Collaboratif sont utilisés pour recommander des produits culturels :

- des films (MovieLens, Ymdb)
- de la musique (Lastfm, Indy)
- des pages web (Del.icio.us, StumbleUpon)

Le Filtrage Collaboratif permet de générer des recommandations personnalisées à l'utilisateur.

Pour les sites commerciaux (Amazon, iTunes, CDnow. . .), il présente donc des enjeux financiers importants.

Filtrage collaboratif: Principes

- Le Filtrage Collaboratif est sous-jacent aux systèmes de recommandation.
- Il regroupe des techniques qui visent à opérer une sélection sur les éléments à présenter aux utilisateurs (Filtrage)
- en se basant sur le comportement et les goûts exprimés de très nombreux utilisateurs (Collaboratif).

Filtrage collaboratif: Principes

Recueil d'information

L'information peut être recueillie de façon:

- Explicite: l'utilisateur attribue des notes aux produits ou des appréciations (*like*)
  - Avantages: Pas d'ambiguïté sur les goûts et les centres d'intérêts de l'utilisateur
  - Inconvénients: biais de déclaration; exagération (souvent constatée)
- Implicite: recueil basé sur le comportement (clics, achats, durée sur une page, ...)
  - Avantages: Objectivité
  - Inconvénients: Grande volumétrie, aucune indication sur l'appréciation

Filtrage collaboratif: Deux approches standards

- Systèmes de filtrage collaboratifs utilisateurs ou approche centrée utilisateurs
  1. Chercher des utilisateurs qui ont les mêmes comportements avec l'utilisateur à qui l'on souhaite faire des recommandations
  2. Utiliser les notes des utilisateurs similaires pour calculer une liste de recommandations pour cet utilisateur.

- Systèmes de filtrage collaboratifs objets ou approche centrée sur les objets

  Approche popularisé par Amazon avec la fonctionnalité "les gens qui ont acheté x ont aussi acheté y".
  - Bâtir une matrice item-item déterminant des relations entre des objets "pairs"
  - Utiliser cette matrice pour proposer des objets.

Filtrage collaboratif: Deux approches standards

Si vous aimez ce film, vous pourriez aimer ...

True Grit

Mort ou vif

Le Bon, la brute et le truand

Fureur Apache

Les clients ayant acheté cet article ont également acheté

Lenovo G50-70 PC portable 15" Noir (Intel Core i3, 6 Go de RAM, disque dur 1 To, Windows 8.1)

(27)

Lenovo G50-70 59412003 Ordinateur portable 15,6" Noir (Intel Core i3, Disque dur 1To, 4 Go de ...

(27)

Case Logic VNAi215 Sacoche en nylon pour Ordinateur portable 15.6" Noir

(12)

EUR 19,99

Logitech Wireless Mouse M235 Souris optique sans fil 2.4 GHz récepteur sans fil USB mercure

(133)

Un site internet, `www.hotelsDauphine.com`, recueille les avis (suivant une notation allant de 1 à 10) suivants de sept personnes pour huit hôtels de la région parisienne:

	Angel	Eden	Brice	Alex	Mania	Bibi	Gabi
Hôtel Saint-Rémy	7	4	10	6			10
Hôtel Courcelles	7	2	8	8	9	4	
Hôtel Gif		8	2	9	2	8	
Hôtel Bures	9		6		8	10	6
Hôtel Orsay	10	4	10	9		9	8
Hôtel Guichet	3	7	2	9		9	8
Hôtel Lozère	5			8	8	8	10
Hôtel Palaiseau	4	6		4	2	8	

Marie Florence, une canadienne régulière en Ile de France a déjà séjourné dans cinq des huit hôtels ci-dessus et leur a attribué les notes suivantes:

	Marie Florence
Hôtel Saint-Rémy	6
Hôtel Courcelles	
Hôtel Gif	
Hôtel Bures	10
Hôtel Orsay	8
Hôtel Guichet	5
Hôtel Lozère	6
Hôtel Palaiseau	

Quel hôtel recommander à Marie Florence (parmi ceux où elle n'a pas encore résidé)?

Plan

Filtrage collaboratif: Approche centré utilisateur

Elle repose sur

- La prédiction de notes:
  considérer la matrice de notes (utilisateurs $\times$ articles)
- matrice à "trous"
- tâche = prédiction des notes manquantes

Un site internet, `www.hotelsDauphine.com`, recueille les avis (suivant une notation allant de 1 à 10) suivants de sept personnes pour huit hôtels de la région parisienne:

	Angel	Eden	Brice	Alex	Mania	Bibi	Gabi
Hôtel Saint-Rémy	7	4	10	6			10
Hôtel Courcelles	7	2	8	8	9	4	
Hôtel Gif		8	2	9	2	8	
Hôtel Bures	9		6		8	10	6
Hôtel Orsay	10	4	10	9		9	8
Hôtel Guichet	3	7	2	9		9	8
Hôtel Lozère	5			8	8	8	10
Hôtel Palaiseau	4	6		4	2	8	

Marie Florence, une canadienne régulière en Ile de France a déjà séjourné dans cinq des huit hôtels ci-dessus et leur a attribué les notes suivantes:

	Marie Florence
Hôtel Saint-Rémy	6
Hôtel Courcelles	
Hôtel Gif	
Hôtel Bures	10
Hôtel Orsay	8
Hôtel Guichet	5
Hôtel Lozère	6
Hôtel Palaiseau	

Quel hôtel recommander à Marie Florence (parmi ceux où elle n'a pas encore résidé)?

- Pour faire une recommandation à Marie Florence en tenant compte des avis des internautes ci-dessus, nous avons besoin d'un mécanisme permettant de déterminer celles ayant des goûts similaires.

- Pour cela, nous allons comparer chaque personne à toutes les autres en calculant un "score de similarité" ou "score de similitude".

- Pour calculer simplement un score de similarité, nous utiliserons la distance de Manhattan ou la distance euclidienne.

  Ainsi, si n représente le nombre d'hôtels pour lesquels les internautes x et y ont attribué une note, alors le score de similarité entre x et y sera assimilé à:

  - leur distance de Manhattan $d(x, y)$ définie par la formule

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

  - leur distance euclidienne $d(x, y)$ définie par la formule

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

  où $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$ sont les vecteurs notes des n hôtels pour lesquels x et y ont attribué une note (les hôtels pour lesquels aucune note n'a été attribuée ne seront pas pris en compte dans cette formule).

- La procédure qui déterminera les recommandations à faire à Marie Florence se déroulera en deux étapes. Pour la décrire, notons pour un hôtel a donné, $\mathcal{C}(a)$ la liste des internautes ayant attribués une note à a, et $x(a)$ la note attribuée à a par l'internaute x.

1. **Étape 1**: Pour chaque hôtel a non visité par Marie Florence, calculer les quantités

$$total(a) = \sum_{x \in \mathcal{C}(a)} \frac{1}{1 + d(x, Marie\ Florence)} \times x(a)$$

$$s(a) = \sum_{x \in \mathcal{C}(a)} \frac{1}{1 + d(x, Marie\ Florence)}$$

$$s'(a) = \frac{total(a)}{s(a)}$$

   La quantité $s'(a)$ permettra de conclure qu'une personne similaire à Marie Florence contribuera de manière plus importante au score global qu'une personne qui lui est différente.

2. **Étape 2**: L'hôtel à recommander à Marie Florence sera l'hôtel qui aura la plus grande somme $s'(a)$.

Autres mesures de similarité: le Coefficient de corrélation de Pearson

Il donne généralement de meilleures résultats que la distance euclidienne lorsque les données ne sont pas parfaitement normalisées, par exemple lorsque les critiques des hôtels sont plus sévères que la moyenne.

$$p(x,y) = \frac{(\sum_{i=1}^{n} x_i\ y_i) - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \times \sqrt{\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}}}$$

où $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$ sont les vecteurs notes des n hôtels pour lesquels x et y ont attribué une note.

Fonction donnant une similarité croissante.

Autres mesures de similarité: la Similarité Cosinus

$$cos(x, y) = \frac{\sum_{i=1}^{n} x_i \; y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}}$$

où $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$ sont les vecteurs notes des n hôtels pour lesquels x et y ont attribué une note.

Quelles mesures de similarité utilisée ?

- Si les échelles de notation sont différentes: coefficient de corrélation de Pearson
- S'il n' y a presque pas de données manquantes: Distance euclidienne
- Si les données sont clairsemées: Similarité cosinus