Algorithmes de recommandation, Cours Master 2, février 2011

Michel Habib
habib@liafa.jussieu.fr
http://www.liafa.jussieu.fr/~habib

février 2011

Plan

1. Recommander un nouvel ami (ex : Facebook)
2. Recommander une nouvelle relation dans un réseau social (ex : Linkdin)
3. Recommander un autre achat dans un logiciel de vente en ligne (ex : Amazon)
4. Recommander une autre vidéo dans un logiciel de recherche de vidéos
5. Choisir ce que l'on vous recommande de lire /voir chez vos amis

Le dernier item est un peu différent :

Sur mon Mur

L'affichage ordonné sur votre mur, des événements les plus pertinents pour vous qui se sont déroulés
dans votre groupe d'amis (Facebook)

Affinités

Site de rencontre par affinités. Recherche de profils proches.

Importance économique du sujet

Annonce récente de thèse 01/02/2010

Nous souhaiterions recruter des étudiants doctorants dans le cadre d'une convention CIFRE afin de développer un algorithme de matching entre les offres d'emploi présentes sur Internet et les données ou profils utilisateurs contenus sur les réseaux sociaux LinkedIn et Facebook.
Plus particulièrement l'algorithme aura pour but de :
Proposer les offres d'emploi les plus pertinentes en fonction du profil d'un utilisateur,
Suggérer des contacts susceptibles d'être intéressés par une offre d'emploi donnée,
Optimiser la création de campagnes publicitaires ciblées sur le réseau Facebook.

À partir de quelles données travaillent ces algorithmes ?

On parle aussi de moteur de recommandation
Un ou plusieurs graphes, plus des données textuelles, des graphes conceptuels ou réseaux sémantiques.

Recommandation d'un nouvel ami

On cherche les sommets du graphe du réseau social ayant le plus de voisins commun avec un sommet x donné.
Recherche de voisinage dans un graphe.

Variante

On cherche suivant le profil : a étudié à Paris Diderot . . .

2 heuristiques possibles

Parmi ceux qui ont le plus de voisins communs, classer par ordre de proximité au profil
Parmi ceux qui sont proches, classer en fonction du nombre de voisins communs

Recommandation d'un livre

On considère le graphe biparti Client – Livres.
Etant donné un client x un livre acheté y , on cherche un biparti maximal complet contenant l'arête xy dans le graphe des achats.
Ce biparti peut éventuellement avoir été précalculé

Variante

On peut associer à chaque livre un graphe conceptuel et rechercher des livres ayant des graphes conceptuels voisins.

One million dollars program

Il y a deux ans, Netflix, le loueur de DVD en ligne américain, avait lancé un concours pour améliorer la pertinence de son moteur de recommandation de films de 10%. Plusieurs équipes de recherche s'étaient lancées dans le défi, mais les propositions avaient du mal à améliorer le moteur de plus de 8,5% .
Le concours est toujours ouvert et les équipes se sont même mises à collaborer entre elles pour essayer de décrocher le gros lot (1 million de dollars de récompense).

Les moteurs de recommandation fonctionnent souvent de la même façon et se contentent d'offrir aux utilisateurs un système de notation pour faire des recommandations adaptées aux notes attribuées. Pour Alex Iskold, il faut distinguer les recommandations personnalisées (adaptées à nos comportements passés), sociales (adaptées au comportement d'utilisateurs similaires) ou sur l'objet.

Données de base

id
user_id
movie_id
rating
timestamp

Les recommandations sociales s'appuient sur un filtrage collaboratif : les gens qui aiment le Seigneur des anneaux vont apprécier Eragon et les Chroniques de Narnia. Le problème de ce type d'approche est que les goûts des gens ne se superposent pas toujours à des catégories aussi simples. Si deux personnes aiment les films de ce type, cela ne veut pas dire qu'ils aimeront les mêmes drames ou les mêmes polars. De même, on pourrait ajouter que si vous aimez un titre de ce genre pour ses qualités de réalisation, il n'est pas sûr que vous apprécierez un autre titre de ce genre dont la réalisation ou l'approche scénaristique seront différentes.

Social information filtering : "word-of-mouth"

- Modélisation du bouche-à-oreille
- Nécessité d'utiliser des modèles issus de la psychologie (car aimer ou ne pas aimer un film, un vin . . .)
- Nécessité de modéliser les cultures.

Première idée d'algorithme

Le principe est pas compliqué, il faut tenter de trouver des utilisateurs similaires (qui tendent à avoir les mêmes gouts (donné le même genre de notes aux films)), pour faire ça comparer chaque personne avec les autres et calculer un score de similitude, pour ce faire on peut soit de calculer une distance euclidienne, soit en utilisant un coefficient de corrélation linéaire.
Une fois que cette méthode est implémentée, il faut l'utiliser pour calculer la personne qui a le meilleur score pour une personne donnée, donc par exemple tenter de trouver la personne qui a donné les scores les plus similaires sur les films qu'elles ont toutes les deux vus. Avec ces deux étapes on se retrouve donc avec la personne ayant noté le plus comme la personne initiale, mais il faut encore trouver un moyen de savoir quels films cette personne à vu et que la personne initiale serait le plus susceptible d'aimer. Pour faire ça il te reste à calculer un score pondéré des notes des autres utilisateurs.

Admettons que Jérôme ai vu les films A, B, C, et D, ainsi que Jessica, Emmanuel, Damien, et d'autres sauf qu'eux en ont également vus d'autres et que justement on voudrait savoir quels films parmi ceux qu'ils ont vu Jérôme devrait regarder aussi). On a déjà calculé des scores de similarité (S1, S2, ..., SN) entre Jérôme et les autres. Alors pour chaque films qu'ils ont vu et que Jérôme n'a pas vu, on multiplie leur note par le score de similitude, et on divise le tout par la somme des score de similarité entre Jérôme et toutes les personnes qui ont vu ce film (afin d'éviter qu'un film vu par plus de personne qu'un autre n'ait forcément un meilleur score). Et voila, il n'y a plus qu'à trier ces résultats de façon décroissante, et nous avons une recommandation des films que Jérôme est le plus susceptible d'aimer au vu des utilisateurs les plus proches de lui question goût !

Etude de cas

Scenario 1

small town bookstore is designing a recommender system to provide targeted personalization for its customers. Transactional data, from purchases cataloged over three years, is available. The store is not interested in providing specific recommendations of books, but is keen on using its system as a means of introducing customers to one another and encouraging them to form reading groups. It would like to bring sufficiently concerted groups of people together, based on commonality of interests. Too many people in a group would imply a diffusion of interests ; modeling reading habits too narrowly might imply that some people cannot be matched with anybody. How can the store relate commonality of interests to the sizes of clusters of people that are brought together ?

Scenario 2

An e-commerce site, specializing in books, CDs, and movie videos, is installing a recommendation service. The designers are acutely aware that people buy and rate their different categories of products in qualitatively different ways. For example, movie ratings follow a hits-bus distribution : some people (the bus) see rate almost all movies, and some movies (the hits) are seen/rated by almost all people. Music CD ratings are known to be more clustered, with hits-bus distributions visible only within specific genres (like "western classical"). Connections between different genres are often weak, compared to connections within a genre. How can the designers reason about and visualize the structure of these diverse recommendation spaces, to allow them to custom-build their recommendation algorithms ?

Scenario 3

An online financial firm is investing in a recommendation service and is requiring each of its members to rate at least products of their own choice to ensure that there are enough overlaps among ratings. The company's research indicates that people's ratings typically follow power-law distributions. Furthermore, the company's marketers have decided that recommendations of ratings can be explainably transferred from one person to another if they have at least 6 ratings in common. Given these statistics and design constraints, what value of should be set by the company to ensure that every person (and every product) is reachable by its recommendation service ?