Gautier Dran
Clément Spies

Les moutons noirs : Etude de l'atypisme

Encadrants : Armelle Brun
Benjamin Gras

Contexte

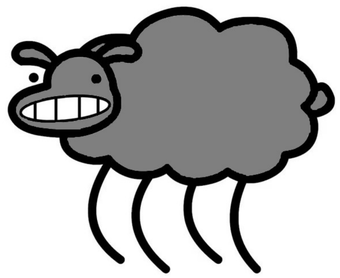
Ce projet est proposé par l'équipe de recherche Kiwi (Knowledge, Information & Web Intelligence) du Loria. Cette équipe a pour but d'améliorer la qualité du service rendu par les systèmes informatiques aux utilisateurs et de faciliter les interactions entre les utilisateurs et les systèmes de recherche. Pour ce faire ils utilisent des systèmes de recommandations.

Présentation du projet

La recommandation dite sociale se sert des avis des utilisateurs qui vous ressemblent, donc qui partagent des préférences avec vous, pour déduire votre avis sur une ressource qui vous sera éventuellement recommandée par la suite. Cependant, chaque être humain possède au moins une préférence qui le différencie des autres et les systèmes de recommandation ne prennent pas en compte de manière particulière ces spécificités. Alors qu'en est-il si vous ne partagez pas assez de caractéristiques avec les autres, que vous êtes trop différent ? Et si vous étiez un mouton noir ? Il s'agit là du problème des moutons noirs de la recommandation sociale. Les systèmes recommandation actuels fournissent des recommandations de mauvaise qualité aux utilisateurs « moutons noirs ». Cependant, tout un chacun a le droit de recevoir des recommandations de qualité, quelles que soient ses spécificités. Il reste donc de nombreuses pistes d'amélioration quant à la détection des ces utilisateurs ainsi que les solutions proposées pour améliorer les recommandations qui leur sont fournies.

Les systèmes de recommandations

Les systèmes de recommandations sont une forme spécifique de filtrage de l'information visant à présenter les éléments d'information qui pourraient intéresser les utilisateurs.
Les cadres d'application de ces systèmes sont multiples : réseaux socio-numériques, marketing digital avec la relation client pour la vente en ligne ou services personnalisés liés à une offre culturelle. Les systèmes de recommandations cherchent donc à prédire la valorisation qu'un utilisateur peut attribuer à un objet (livre, musique, film, etc) ou un élément social (personne, groupe, etc), c'est le cas pour les sites tels que Deezer (musiques), Amazon (livres, films), Facebook (personnes, groupes) et pleins d'autres. Les SR ont comme intérêt pour l'utilisateur de réduire le temps de recherche d'informations et découvrir des produits difficiles à trouver ou qui pourraient l'intéresser. Et ils ont un intérêt pour le fournisseur du service : celui d'orienter le client et d'augmenter ses bénéfices.



Travail réalisé

Notre objectif était de permettre de faire des recommandations de films aux utilisateurs et plus particulièrement aux utilisateurs atypiques (moutons gris). Nous avions donc choisis de classifier nos données par rapport aux films. En effet, regrouper les films nous permis d'identifier des groupes de films similaires suivant différentes critères. Grâce à cela nous devions être en mesure d'améliorer les recommandations quant aux moutons gris. Nous disposions d'un fichier de données de 100 000 lignes. Chaque ligne disposant de 3 éléments : l'identifiant de l'utilisateur, l'identifiant du film et la note (de 1 à 5) que l'utilisateur a attribuée au film. Sur ces 100 000 lignes nous avions 1682 films ainsi que 943 utilisateurs.

Pour traiter ces films il a fallu choisir des méthodes de clustering. Dans un premier temps, nous nous sommes penchés sur une liste d'algorithmes : l'algorithme de maximisation de l'espérance, les cartes auto-organisatrices, la Classification hiérarchique ascendante, le k-means, le k-médoïdes, DBSCAN et OPTICS.

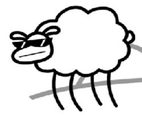
Les premiers algorithmes que nous avions "validés" sont le K-means et le K-médoïdes. Nous avions, pour commencer, utilisé l'agorithme du K-Means sur un fichier de 100 données que nous avions créé à partir du fichier des 100 000 données. Cela nous a permis de comprendre le fonctionnement et d'analyser les résultats d'un premier algorithme de clustering.

Une fois que les films étaient regroupés en cluster, nous avions émis 5 hypothèses :
- Les films avec une forte erreur de recommandation sont regroupés dans un ou deux clusters.
- Les films avec une faible erreur de recommandation sont regroupés dans un ou deux clusters.
- Les films très controversés sont réunis dans un cluster.
- Les films très controversés sont répartis de manière égale dans les clusters.
- Il existe un cluster regroupant des films qui plaisent aux utilisateurs atypiques

Nous avons choisis de nous intéresser plus particulièrement aux deux premières hypothèses. Car dans le cas des utilisateurs atypiques, nous avions très peu de données par rapport à l'erreur de recommandation.

Résultats

Test de l'hypothèse : "Les films avec une forte erreur de recommandation sont regroupés dans un ou deux cluster". Tout d'abord, il fallait définir ce que nous appelions une forte erreur de recommandation. Nous avions donc regarder les données que nous avions par rapport à cette erreur. Pour cela nous avions utilisé une boite à moustache afin de voir comment étaient réparties nos données. Et ensuite en regardant leurs répartitions dans les clusters nous pouvions voir que les films avec un fort taux d'erreur moyen étaient répartis dans le même cluster. Notre hypothèse était donc valide. Nous avions effectué le même principe pour l'hypothèse : "Les films avec une faible erreur de recommandation sont regroupés dans un ou deux clusters" et avions découvert que l'hypothèse était aussi valide.
On peut voir qu'avec les configurations citées précédemment, il est possible de valider nos hypothèses. Nous pourrions donc affirmer qu'il est possible grâce à une classification des films d'identifier les films sur lesquels nous avons un faible/fort taux d'erreur de recommandation. Nous pourrions aussi affirmer qu'il est possible, grâce à une classification des films, d'identifier les films sur lesquels nous sommes sûrs d'avoir assez d'informations afin de les recommander avec précision et les films sur lesquels nous avons besoin de plus d'informations car actuellement les données ne permettent pas d'avoir une bonne recommandation.



Le clustering

Le principe du clustering est de diviser automatiquement un ensemble de données en différents groupes, appelés Clusters. Le regroupement des données se fait par mesure de distance. Chaque donnée est représentée par des caractéristiques qui sont utilisées pour calculer la distance entre deux éléments. Plus ces distances sont faibles plus les éléments seront proches.
Pour que le clustering soit optimal il faut que l'inertie intra-classe (c'est-à-dire les distances entre les points et le centre du cluster) soit minimisée, et que l'inertie inter-classe (c'est à dire les distances entre chaque cluster) soit maximisée.
Le domaine d'applications du clustering est vaste : Reconnaissances des formes, analyse de données spatiales, traitement d'images, recherche d'informations ou encore web mining.

UNIVERSITÉ DE LORRAINE

UFR ATHÉMATIQUES ET INFORMATIQUE
Penser l'informatique autrement

Loria
Laboratoire lorrain de recherche en informatique et ses applications

KIWI
KNOWLEDGE, INFORMATION and WEB INTELLIGENCE