



Un système de recommandation basé sur l'analyse statistique implicative

Raphaël Couturier*

*Laboratoire d'Informatique de l'université de Franche-Comté (LIFC)
 IUT de Belfort-Montbéliard, BP 527, 90016 Belfort, France
 Raphael.couturier@iut-bm.univ-fcomte.fr

Résumé. Ce travail a pour objectif de proposer une nouvelle méthodologie afin de construire un système de recommandation. Notre approche repose sur l'ASI et ne nécessite pas d'avoir un profil d'utilisateur afin d'effectuer des recommandations. Afin de valider notre démarche, nous avons appliqué notre système sur les critiques de la presse du site d'« allociné » qui rassemble les critiques des films récents.

1 Introduction

L'analyse statistique implicative (ASI) a été développée par Régis Gras et ses collaborateurs. Elle permet d'établir des règles d'association à partir d'un ensemble de données croisant sujets et variables. Pour de plus amples informations le lecteur intéressé est invité à consulter (Gras et al. 1996, Gras et al. 2004). La mesure la plus utilisée afin d'établir une règle d'association est la probabilité conditionnelle. L'algorithme *a priori* permet de calculer efficacement les règles d'association avec une telle mesure (Agrawal et al. 1993). Les avantages de l'ASI par rapport à la probabilité conditionnelle peuvent être résumés par :

- l'implication n'est pas linéaire en fonction du nombre de contre-exemples,
- elle est résistante au bruit,
- elle rejette les règles triviales, c'est-à-dire celles qui n'apportent pas de nouvelles connaissances,
- elle tient compte de la taille des effectifs,
- elle est utilisable sur de nombreux types de variables (binaires, fréquentielles, modales, intervalles, ...).

Actuellement trois types de représentation sont proposées au sein de l'outil CHIC présenté dans (Couturier et Gras 2005) : un arbre des similarités non orienté, un arbre cohésitif orienté et un graphe implicatif.

Avec le développement d'internet, des sites permettant aux visiteurs de noter ou critiquer des produits ou des éléments ont vu le jour. En général, une personne donne son avis sur un produit par l'intermédiaire d'une note et d'un commentaire. L'analyse de données textuelles bien qu'étant en pleine expansion, se heurte encore à des difficultés. Par contre, il est possible d'utiliser les notes des utilisateurs afin d'établir les tendances des avis et ainsi de proposer un système de recommandation. Pour un élément ou produit donné, on peut ainsi proposer d'autres éléments ou produits en fonction des avis recueillis. Notre objectif dans ce papier est de décrire comment réaliser un tel système de recommandation basé sur l'ASI.

Dans la suite, nous présentons notre approche et une application sur les critiques de la presse concernant les films récents projetés au cinéma (ces données ont été recueillies sur le site www.allocine.fr).

2 Méthodologie de l'approche choisie

Avant de détailler l'approche que nous avons choisie, nous présentons le contexte que nous avons retenu. Il nous semble nécessaire que les personnes qui critiquent puissent pouvoir s'identifier afin d'éviter qu'une personne se connecte avec des noms d'utilisateur différents suivant les sessions. Si le nombre d'éléments à



noter est important, il paraît indispensable que les utilisateurs donnent leur avis sur de nombreux produits. Une solution consiste à inciter les utilisateurs à se prononcer à la faveur d'une récompense quelconque.

La prise en compte des avis des utilisateurs est possible selon diverses méthodes. Supposons que les utilisateurs puissent donner une note entre 1 et 10 , voici diverses méthodes envisageables:

- Il est possible de ramener ces valeurs entre 0 et 1 en soustrayant 1 à toutes les valeurs et en divisant ensuite le résultat par 9 .
- Il est possible de ramener les valeurs entre 0.1 et 1 en divisant les notes par 10 .
- Il est concevable de penser que les personnes qui mettent une note supérieure à la moyenne jugent plutôt bien l'élément et celles qui mettent une note inférieure n'ont pas apprécié celui-ci. Dans ce cas, on peut ramener les notes comprises dans l'intervalle $[5,10]$ dans l'intervalle $[0,1]$ en prenant en compte que l'avis est positif et ramener les notes comprises dans l'intervalle $[1,4]$ dans l'intervalle $[0,1]$ en indiquant que l'avis est négatif (la note 1 correspond à 1 et la note 4 correspond par exemple à 0.25).

D'autres échelles de valeur peuvent présenter d'autres avantages.

Après avoir formaté les données, nous utilisons le moteur de CHIC qui permet d'établir les règles d'associations entre les variables. Dans notre cas, les règles nous indiquent l'avis général des utilisateurs sur les produits. Selon que l'on considère des avis positifs uniquement ou positifs et négatifs, les résultats seront différents. Il est d'ailleurs possible de mélanger les avis positifs et négatifs. Le choix de la formule permettant de prendre en compte les avis positifs et négatifs dépend du domaine considéré et un expert des données sera plus apte à donner un avis motivé.

3 Travaux relatifs

Il existe de nombreux travaux sur ce sujet. Par exemple dans (Breese et al. 1998) les auteurs proposent deux méthodes de recommandation en fonction des produits qu'un utilisateur a déjà critiqués. Ces approches sont différentes de celle que nous proposons. Elles utilisent les critiques d'un utilisateur afin de le recommander. L'avantage est de pouvoir faire une recommandation peut être plus fine mais l'utilisateur est obligé de se "dévoiler" avant d'avoir des recommandations. L'algorithme utilisé pour les deux approches de ce papier repose sur une similarité entre l'utilisateur qui souhaite être recommandé et les autres utilisateurs de la base de données. L'algorithme repose sur la technique de filtrage collaboratif.

Dans (Rashid et al. 2005), les auteurs proposent un algorithme basé sur la technique des k-voisins afin d'établir une recommandation pour un utilisateur ayant déjà un profil, c'est-à-dire ayant déjà critiqué lui-même des produits.

Dans (Shahabi et Chen 2003), les auteurs mélangent la technique de filtrage collaboratif avec un algorithme génétique.

La contrainte pour un utilisateur de devoir noter certains produits avant d'avoir des recommandations peut se révéler gênante, c'est pourquoi dans notre approche nous préférons faire des recommandations que l'on peut qualifier d'anonyme.

4 Application aux critiques de la presse sur des films récemment sortis au cinéma

Nous avons choisi de recueillir les données présentes sur le site internet d'« allociné » en ne prenant en compte que l'avis de la presse. Ceci est justifié par le fait que la presse critique de nombreux films (sans forcément les critiquer tous systématiquement). De nombreux cinéphiles critiquent également les films mais un grand nombre d'entre eux ne le font pas de façon régulière.



Afin de pouvoir appliquer notre approche sur différents sites, nous avons développé des scripts qui rapatrient les pages contenant des critiques. Selon la mise en forme des données, il est bien évidemment indispensable de modifier ces scripts. Néanmoins, certaines parties comportent de fortes similitudes.

Les données recueillies sont placées dans une base de données. Actuellement, la base comporte 16657 avis sur un total 1248 films critiqués par 49 revues de presse. Les notes sur le site d'« allociné » varient de 1 à 4 et sont représentées par des étoiles. Nous avons choisi de diviser la note par 4. À partir de ces données, le moteur de CHIC qui calcule les implications avec un algorithme inspiré de celui décrit dans (Agrawal et al. 1993), établit 1225470 règles.

Ces règles nous informent que la presse qui a apprécié un film a également apprécié une liste d'autres films. Ainsi l'utilisateur, suivant ces goûts, pourra suivre les recommandations de la liste s'il a apprécié le film sélectionné. Au contraire il préférera peut-être ne pas aller voir tel autre film proposé dans la liste car il n'a pas jugé bon le film à la base de la recommandation.

Dans le tableau TAB 1, nous avons synthétisé le nombre de règles supérieur à un certain seuil pour les trois types d'indices que peut calculer le moteur de CHIC.

| seuil | Implication classique | Implication entropique | Probabilité conditionnelle |
|-------|-----------------------|------------------------|----------------------------|
| 0 | 1225470 | 214633 | 1225470 |
| 10 | 1225470 | 214633 | 1095122 |
| 20 | 1225467 | 214633 | 876017 |
| 30 | 1225467 | 213896 | 623843 |
| 40 | 1224273 | 202400 | 402896 |
| 50 | 1194931 | 176142 | 214633 |
| 60 | 977090 | 136874 | 116703 |
| 70 | 675828 | 91557 | 57398 |
| 80 | 370491 | 39747 | 18367 |
| 90 | 119117 | 12064 | 7467 |
| 95 | 35668 | 2318 | 5698 |

TAB 1 – Nombre de règles obtenues en fonction de certains seuils avec les trois indices.

Au vu de ces résultats, il nous semble intéressant de prendre l'implication entropique afin d'établir les recommandations pour notre système. En effet, cet indice est plus discriminatif que les deux autres et en plus il présente des caractéristiques plus intéressantes (Gras et al. 2004).

Afin de permettre aux utilisateurs de naviguer dans les recommandations, nous avons disposé les résultats des règles sur un site internet. La figure FIG 1 illustre une capture d'écran de l'interface du site. Au départ l'utilisateur rentre une partie du nom d'un film. Les films comportant le mot en question apparaissent et l'utilisateur peut cliquer sur le film de son choix. Ensuite les recommandations du film en fonction de la presse apparaissent sur la partie gauche de la figure. Sur la capture d'écran, nous avons choisi de montrer uniquement les 20 premières recommandations. Celles-ci sont classées par ordre décroissant en fonction de l'intensité d'implication entropique qui est indiqué à droite de chaque film à titre informatif.



L'Un reste, l'autre part - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://localhost/demo_allocine/?item_id=55904

The Mozilla Orga... Latest Builds

Nom du film Chercher

L'Un reste, l'autre part [IMDb](#) [Allocine](#)

Recommandations :

| | |
|-------------------------------------------------|------|
| A tout de suite | 87.7 |
| Rois et reine | 86.1 |
| Le Château ambulant | 84.4 |
| Vera Drake | 84.1 |
| Mondovino | 76.2 |
| Sideways | 73.7 |
| Le Couperet | 73.6 |
| Maria, pleine de grâce | 72.9 |
| Le Cauchemar de Darwin | 70.6 |
| Aviator | 69.7 |
| La Marche de l'empereur | 66.2 |
| La Chute | 65.8 |
| Les Temps qui changent | 64.3 |
| Le Promeneur du Champ de Mars | 64.3 |
| Mar adentro | 64.3 |
| 2046 | 62.5 |
| Ray | 61.5 |
| Le Secret des poignards volants | 61.3 |
| Prendre femme | 61.3 |
| Terre promise | 58.3 |

Notes :

| | |
|------------------------------------|--|
| Elle | |
| Score | |
| Première | |
| Paris Match | |
| MCinéma.com | |
| France Soir | |
| Studio Magazine | |
| Le Figaro | |
| Ouest France | |
| Libération | |
| Le Point | |
| Ciné Live | |
| Le Figaroscope | |
| Zurban | |
| L'Humanité | |
| TéléCinéObs | |
| Cinéastes | |
| Rolling Stone | |
| Le Monde | |
| Cahiers du Cinéma | |
| Télérama | |
| L'Express | |
| Télé 7 Jours | |
| Les Inrockuptibles | |
| aVoir-aLire.com | |
| Positif | |

Done Adblock

FIG 1 – Capture écran d'une recommandation



L'utilisateur peut cliquer sur les deux liens l'informant sur le film en question. Le premier se réfère à la base IMDB qui contient une liste mondiale des films de cinéma. Le second lien contient la page sur le film choisi sur le site d' « allociné ». La partie droite de l'écran indique les notes de la presse qui ont émis une critique sur le site en question, les notes varient entre 1 et 4 et sont représentées par des étoiles. Ainsi avec l'exemple de figure, trois revues de presse ont vraiment apprécié le film « L'un reste, l'autre part » et trois autres revues ne l'ont vraiment pas apprécié.

| film | Recommandations (indice d'implication entropique) |
|-------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Mar Adentro | Vera Drake (89.6) Rois et reine (88) Le Château ambulant (86.6) A tout de suite (84.2) Sideways (81.7) Le Couperet (78.9) Le Cauchemar de Darwin (78.6) Mondovino (78.1) Maria, pleine de grâce (75.3) |
| Les indestructibles | Le Château ambulant (81.8) Rois et reine (78.9) Nobody knows (75.3) A tout de suite (75) Collateral (74.6) Spider-Man 2 (74.3) Clean (74.1) Memories of murder (73.1) |
| La Demoiselle d'Honneur | A tout de suite (95.7) Rois et reine (90.5) Clean (84.7) Spider-Man 2 (83.8) Mondovino (83.1) Le Château ambulant (82.3) Memories of murder (80) Les Indestructibles (79.3) |
| 36 quai des orfèvres | Le Château ambulant (77.3) Le Couperet (76.5) A tout de suite (76.3) Rois et reine (74.1) Aviator (73.2) Maria pleine de grâce (72.6) Vera Drake (70.8) Le Cauchemar de Darwin (68.7) |
| 21 grammes | Buongiorno, notte (96.3) S21, la machine de mort Khmère rouge (93.2) Saraband (91.7) Clean (91.7) Rois et reine (91.7) A tout de suite (90.9) Triple agent (88.8) Le Retour (88.7) Mondovino (88.5) Open range (87.7) |
| Osama | Feux rouges (91.6) Buongiorno, notte (89.2) S21, la machine de mort Khmère rouge (88.6) A tout de suite (88.3) Nobody knows (87.7) Rois et reine (87.3) Clean (85.8) Gerry (84.7) |
| L'Esquive | A tout de suite (92.1) Lost in translation (87.9) Rois et reine (87.2) Buongiorno, notte (86.6) Saraband (82.3) Memories of murder (81.7) S21, la machine de mort Khmère rouge (77.9) Spider-Man 2 (77.8) |

TAB 2 – Quelques exemples de recommandations proposés.

Dans le tableau TAB 2 nous avons relevé quelques propositions. Pour chaque film proposé, nous faisons apparaître l'indice d'implication entropique. Par exemple, si un utilisateur a bien apprécié « La demoiselle d'honneur », alors en fonction de critiques de la presse, on lui recommandera d'aller voir « A tout de suite » (avec une intensité d'implication égale à 0.957), d'aller voir « Rois et reine » (avec une intensité d'implication égale à 0.905) etc.

De toute évidence les films que toute la presse a plébiscités se retrouvent dans plusieurs recommandations. Il est tout à fait envisageable d'affiner les recommandations en proposant différentes catégories de listes de films, par exemple les films à découvrir, les grands succès, les nominés, etc. Par ailleurs, il semble également intéressant de faire des propositions en fonction de la catégorie du film, c'est-à-dire que les recommandations d'un film d'action seront du même genre et ainsi de suite.



5 Conclusion

Dans cet article nous avons présenté comment mettre en place un système de recommandation basé sur l'implication statistique, entropique plus précisément. L'avantage est de pouvoir établir des recommandations sans avoir un profil utilisateur. Certains préféreront employer des algorithmes pour lesquels l'utilisation d'un profil est indispensable pour engendrer des recommandations. Il paraît difficile de comparer notre approche à d'autres approches puisqu'elles ne reposent pas sur les mêmes principes et ne font pas les mêmes hypothèses.

Nous avons appliqué notre technique sur les critiques de la presse du site « allociné » qui recense les films récents projetés au cinéma. Ainsi, notre système propose des recommandations aux personnes qui ont apprécié un film, cette recommandation est basée sur les critiques de la presse.

Références

- Agrawal R., Imielinski T. et Swami A. (1993), Mining association rules between sets of items in large databases. Proceedings of the International Conference on Management of Data, 1993, pp 207-216.
- Breese J.S., Heckerman D. et Kadie C. (1998), Empirical analysis of predictive algorithms for collaborative. Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, pp 43-52.
- Couturier R. et Gras R. (2005), Chic : traitement de données avec l'analyse implicative. Actes d'Extraction et gestion des connaissances (EGC'2005), Vol. 2, pp 679-684.
- Gras R., Ag Almouloud S., Bailleul M., Lahrer A., Polo M., Ratsimba-Rajohn H. et Totohasina A. (1996), L'implication Statistique, La Pensée Sauvage.
- Gras R., Couturier R., Blanchard J., Briand H., Kuntz P. et Peter P. (2004), Quelques critères pour une mesure de qualité de règles d'association. Un exemple : l'implication statistique. Mesures de qualité pour la fouille de données}, RNTI-E-1, Cepaduès Editions, pp 3-32.
- Rashid A.M., Karypis G. et Riedl J. (2005), Influence in ratings-based recommender systems: An algorithm-independent approach. Proceedings of the SIAM International Conference on Data Mining, À paraître.
- Shahabi C. et Chen Y.S. (2003), An adaptative recommendation system without explicit acquisition of user relevance feedback. Distributed and Parallel Databases, Vol. 14, pp 173-192.

Summary

In this work we proposed a new methodology to make a recommendation system. It is based on the ASI (Implicative Statistical Analysis). Our approach does not require to profil users before recommending them. Some experiments are reported on the press critics from the allocine website which provides information of recent movies.