Analyse de réseaux sociaux et Recommandation de contenus non populaires

Cécile Bothorel*

*TELECOM Bretagne, Dép. LUSSI, UMR CNRS 3192 Lab-STICC,
Technopôle Brest-Iroise CS 83818, F-29238 Brest Cedex 3
cecile.bothorel@telecom-bretagne.eu
http://cecile-bothorel.labocommunicant.net/

Résumé. Nous présentons une méthode adressant le problème de la recommandation dans la Longue Traîne, et plus généralement celui du démarrage à froid. Les contenus non populaires, peu annotés, sont difficiles à recommander. Notre originalité est duale et repose sur le fait de capturer la richesse d'annotations du Web Social d'une part, et d'autre part, d'exploiter le fait que les internautes, via un réseau social, sélectionnent eux-mêmes leurs prescripteurs de contenus. La méthode *Social Popularity* détecte des communautés dans un réseau social de fans de cinémas, puis calcule, contextuellement à ces communautés, des similarités entre films rares. La méthode montre des résultats préliminaires intéressants, elle permet d'augmenter notablement le taux de rappel en retrouvant d'avantage de films sélectionnés par les utilisateurs (les vrais positifs). La précision reste gloablement faible comme les autres méthodes testées, ce qui montre qu'il est très difficile de diminuer le nombre de prédictions fausses.

1 Introduction

La recommandation est un domaine scientifique qui cherche à personnaliser l'accès à l'information pour un utilisateur donné, et ainsi lui faciliter le choix de contenus dans un catalogue trop vaste pour qu'il puisse s'en faire une idée d'ensemble. En pratique, les systèmes de recommandation, à partir de connaissances sur un utilisateur, filtrent un ensemble de contenus et produisent une liste, souvent ordonnée, de ces quelques contenus sélectionnés et jugés pertinents pour lui.

Les travaux de recherche se concentrent sur la conception d'algorithmes de sélection de contenus pour un utilisateur donné dont on connaît un profil de goûts et/ou ses achats. L'exemple emblématique est le site de e-commerce Amazon [1] qui oriente le visiteur vers des contenus que d'autres visiteurs ont appréciés. Des méthodes de recommandation sont utilisées sur d'autres sites Internet de vente par correspondance (Fnac [2], Virgin [3], etc.) ou encore sur les plates-

1. http ://www.amazon.com
2. http ://www.fnac.com
3. http ://www.virgin.com

formes musicales (Lastfm [4], Pandora [5], etc.) ou des sites de vidéo à la demande tels que Netflix [6].

Resnick et Varian (1997) dégagent deux grandes familles : les techniques basées sur le *filtrage collaboratif* qui recherchent des similarités de profils utilisateurs en se basant sur des notations (un nombre d'étoiles, la liste des achats passés) ou bien des similarités de profils de contenus sur la base de descripteurs. La deuxième famille de techniques à base de *filtrage de contenus* recherche une adéquation thématique entre un profil d'utilisateur et des profils des contenus. Candillier et al. (2009) comparent ces différentes méthodes, avec différentes mesures de similarité, ainsi que différentes méthodes d'évaluation sur deux jeux de données de référence, dont l'un très actuel, celui du concours Netflix [7].

L'état de l'art des systèmes de recommandation liste quantité d'algorithmes, de calculs de similarités, de combinaison de techniques. Le concours Netflix a offert un challenge motivant aux chercheurs qui ont fait rapidement progressé l'état de l'art et produit des techniques complexes de qualité. Certains, comme Herlocker et al. (2004), se demandent si les gains en qualité de recommandation sont dorénavant perceptibles par l'utilisateur : il y a en effet une sorte de "barrière magique" qui empêche les systèmes d'atteindre la perfection étant donné que les notations par les utilisateurs eux-mêmes sont inconsistantes voire contradictoires.

Pourtant il reste des verrous scientifiques dans le domaine, dont celui de la recommandation de contenus non populaires.

1.1 Problème de la recommandation de la Longue Traîne

La recommandation est un marché avant tout. Piller (2007) chiffre l'impact de la recommandation en termes de revenus, et cite, parmi d'autres exemples, le cas de CleverSet, un fournisseur de moteur de recommandation pour magasins en ligne, qui annonce un accroissement de revenus de 22% par visiteur en moyenne. En 2006 Amazon estimait à 35% l'augmentation de ventes grâce aux suggestions d'achat contextuelles aux articles visionnés [8]. L'année suivante Netflix lance son concours spectaculaire pour améliorer la qualité de ses recommandations. Même sans beaucoup de références étayées sur le sujet (information sensible ?), on l'aura compris, les enjeux sont colossaux. Il s'agit pour des distributeurs comme Amazon de reconstruire sur Internet un linéaire de produits optimisé parmi plus de 3 millions de produits.

Anderson (2006) introduit le phénomène de la Longue Traîne qui met en évidence des produits de niche, peu populaires, qui, chacun pris séparemment, ne génèrent que peu de revenus, mais qui collectivement représentent un marché non négligeable (Fig. 1 [9]). Les distributeurs de bien culturels en ligne, non tributaires de la contrainte physique de leur étalage à optimiser, peuvent se permettre de proposer les produits de niche.

La disponibilité de contenus en ligne très variés ne veut pas dire pour autant qu'ils soient plus achetés qu'auparavant. Des études montrent que les contenus populaires restent les contenus favoris des acheteurs, qu'ils soient gros ou petits consommateurs. En musique, par exemple,

4. http ://www.lastfm.com
5. http ://www.pandora.com
6. http ://www.netflix.com
7. Compétition lancée par Netflix, Inc. pour trouver un algorithme de recommandations de films surpassant le leur de manière significative : gain de plus de 10% dans la pertinence des notes prédites (mesure RMSE) sur le jeu de test fourni. Netflix a récompensé les gagnants d'une prime d'un million de dollars en 2009.
8. http ://venturebeat.com/2006/12/10/aggregate-knowledge-raises-5m-from-kleiner-on-a-roll/
9. D'après http ://www.cs.cmu.edu/ jhm/Google Knows How to Flirt.htm

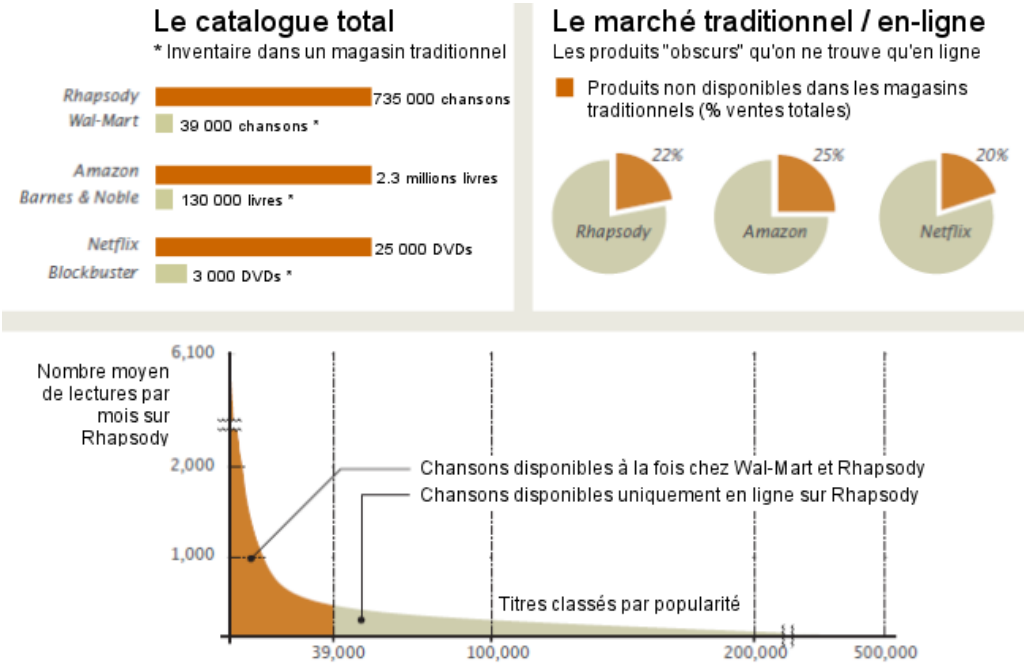


FIG. 1 – *Si l'on compare le marchand de musique en ligne Rhapsody et le magasin traditionnel Wal-Mart, nous voyons que le marchand en ligne offre 20 fois plus de titres, parmi lesquels nous trouvons les produits de la Longue Traîne, très nombreux et moins populaires. 95% du catalogue de Rhapsody ne figure pas chez son concurrent traditionnel et génère 22% des ventes. Les moteurs de recommandation sont un fort enjeu pour mettre en avant ces contenus.*

parmi les 2,4 millions de morceaux numériques vendus en 2007 aux USA (la plupart via iTunes), 24% des morceaux ne se sont vendus qu'une seule fois, et 91% moins de 100 copies.

Cependant les gros consommateurs s'intéressent plus aux contenus moins populaires, ce que Google met également en évidence sur la plateforme YouTube : on pourraît croire que les visiteurs ne recherchent que les vidéos les plus vues et partagées, celles qui font du "buzz". Or d'après Baluja et al. (2008), pour les utilisateurs non occasionnels, ceux qui sont revenus au moins 4 fois sur le site pendant les 92 jours analysés en 2007, les attentes vont au-delà des contenus populaires [10]. Le rôle des systèmes de recommandation est de satisfaire et fidéliser ces gros consommateurs. L'enjeu, bien au-delà de cette fidélisation, est de promouvoir les contenus de niche dont la marge commerciale est plus importante que sur les produits très populaires.

Or, Park et Tuzhilin (2008) montrent qu'il est difficile de prédire une note pour les contenus de la Longue Traîne : en utilisant 9 méthodes d'apprentissage différentes, ils démontrent que

10. Les "grands" consommateurs de VoD chez le loueur Netflix ont le même comportement. Cf. Tom F. Tan and Serguei Netessine. *Is Tom Cruise threatened ? Using Netflix Prize data to examine the long tail of electronic commerce*. Working Paper, 2009

l'erreur de prédiction grandit pour les contenus pour lesquels on ne dispose que de peu de notations. Ils appellent ce phénomène le *problème de la recommandation de la Longue Traîne*. Pour pallier ce problème, ils proposent une méthode consistant à regrouper les contenus de la Longue Traîne en groupes de contenus similaires puis d'appliquer des méthodes prédictives sur les regroupements et non sur chacun des contenus, cumulant ainsi le peu de notations en une masse critique de notations exploitables. Ils proposent également des pistes pour détecter la limite entre la Tête et la Queue de la distribution, ainsi que pour définir le nombre de clusters de contenus rares. Ces deux paramètres sont difficiles à choisir, varient d'un jeu de données à un autre et affectent le résultat de manière significative.

Le *problème de la recommandation de la Longue Traîne* n'a été abordé en tant que tel qu'assez récemment dans le contexte de la recommandation (atelier dédié à la recommandation de musique de la Longue Traîne à la conférence RecSys'10). Néanmoins la problématique des contenus sur lesquels on ne dispose que peu de données d'usage est connue sous le vocable de *démarrage à froid*.

1.2 Le Web pour pallier le manque de données d'usage

Sans données d'usage sur un contenu, la recommandation s'avère risquée : le rapprochement entre les goûts d'un utilisateur et un contenu "peu" voire "pas" connu est hasardeux. Ce problème, bien connu, est particulièrement épineux au lancement d'un service (phase de *démarrage à froid*, ou *cold start* en anglais), et se repose également pour tout nouvel utilisateur ou pour tout nouveau produit.

Notre problématique est proche. Les contenus non populaires sont des contenus pour lesquels un service ne peut recueillir beaucoup d'usage (achats, commentaires, notes) et cela sur le long terme. Le démarrage à froid est en quelque sorte permanent pour ces contenus de niche.

Schein et al. (2002) proposent un tour d'horizon de ce problème. Une solution consiste à utiliser des métadonnées décrivant de manière riche le contenu, et de le recommander aux personnes qui ont déclaré leurs goûts de manières précises (filtrage de contenus) ; sans information de profil utilisateur, il est possible de trouver des similarités de contenus sur la base de ces métadonnées et de proposer des recommandations au regard des annotations des utilisateurs sur des contenus similaires.

Cette approche à base de filtrage de contenus est très coûteuse en terme d'acquisition de métadonnées. La tendance ces dernières années est d'exploiter la richesse du Web et de capturer les métadonnées disponibles sur des sites spécialisés tels que IMDb (Internet Movie Database).

En ce qui concerne les goûts des utilisateurs, une idée similaire est d'alimenter des profils de notation depuis le web communautaire et social et ainsi augmenter le nombre d'individus entrant dans le calcul de similarités d'usage (approche filtrage collaboratif). Un exemple récent est celui qui puise dans les blogs et autres commentaires textuels laissés par des internautes à propos de films des informations d'opinions. Après un travail d'analyse textuelle des opinions (opinion mining), avec mise en évidence des mots porteurs d'opinion permettant de prédire des notes de films (Poirier et al., 2009), l'idée est d'augmenter des matrices d'usage de films d'un site marchand par des profils collectés à partir du web, et ainsi d'alimenter une matrice trop creuse au moment du démarrage du site. Une matrice de similarités de films apprise sur des données externes peut être réutilisée par un moteur de recommandation basé sur un filtrage collaboratif, à condition bien sûr de faire la correspondance entre les films disponibles dans

le catalogue et ceux de la matrice. Poirier et al. (2010) montrent que cette approche est plus intéressante que le filtrage de contenus en terme de qualité de recommendation "à froid". Les usages collectés depuis le site communautaire Flixster [11] peuvent donc se transposer sur un service VoD tel que Netflix.

1.3 Notre contribution

Notre travail s'inscrit dans cette démarche d'exploiter la richesse du Web et en particulier de :

- profiter de la masse de données d'usage accessibles concernant les contenus non populaires ;
- collecter des similarités d'utilisateurs qui ont une forte propension à s'exposer sur le Web et à se regrouper autour de ces contenus non populaires.

En nous basant sur un site à très large audience, tel que Flixster [12], nous collectons une grande masse de données d'annotations de films. Les internautes créent des profils, et publient leurs opinions concernant les films qu'ils ont vus, sous forme de commentaires textuels et de notes. Un tel site est très actif, les internautes y forment des communautés de goûts diversifiés, laissant place aux goûts orignaux ; nous faisons l'hypothèse forte ici de capturer des usages inhérents à la Longue Traîne.

D'autre part, les internautes utilisent des fonctionnalités sociales et sur un site tel que Flixster, en plus d'annoter des films, ils déclarent des listes d'amis, à la manière de Facebook [13], et partagent leurs goûts avec leurs amis. Le choix des amis repose pour une partie sur le profil de cinéfile affiché ; c'est ici une deuxième hypothèse que nous faisons, les internautes se regroupent selon des goûts similaires, ou en tous les cas, recherchent des interactions avec des internautes dont les goûts les intéressent.

Nous cherchons dans ce travail à apprendre des *similarités de films non populaires* en réponse au problème de la recommandation de la Longue Traîne.

Notre contribution se situe à deux niveaux. Dans la méthodologie proposée premièrement, avec l'exploitation du Web Social où le volume de données, y compris sur les films non populaires, reste significatif. Dans la technique mise en œuvre deuxièmement, avec un algorithme de détection de communautés dans les réseaux sociaux qui capture les regroupements d'internautes qui naturellement tissent des liens sociaux avec les cinéphiles qui les intéressent. Les données d'usage utilisées sont de deux types : les notes exprimées sur les films ainsi que les marques sociales d'intérêt pour les publications des pairs. Nous proposons dans ce papier une première ébauche exploitant ces deux types de données de sorte à réduire l'effet Longue Traîne des contenus non populaires.

L'état de l'art de la section suivante dresse un panorama aussi large qu'hétéroclite de l'exploitation du Web Social pour la recommandation de contenus. Nous proposons ensuite une méthode d'analyse d'un réseau social de cinéphiles en vue d'adopter une approche collaborative dans le contexte d'un système de recommandation. L'évaluation de notre proposition montre l'intérêt de l'approche même si nous n'avons pas encore mené toutes les validations attendues, en particulier celle de tester les similarités de films apprises dans le contexte d'un système de recommandation existant tel que Netflix à la manière de Poirier et al. (2010).

11. http ://www.flixster.com
12. Nous utilisons le même jeu de données que Poirier et al. (2010), nos travaux étant complémentaires.
13. http ://www.facebook.com

2 Exploiter le Web Social pour la recommandation

Cette section fait un balayage assez large et très hétéroclite des études liées au Web Social pour la recommandation de contenus. Ce type de travaux est assez récent, il nous paraît important de présenter les différentes orientations très diversifiées avant de détailler à proprement parler notre méthode basée sur l'analyse de réseaux sociaux pour la recommandation. Les parties de cette section n'ont pas de lien entre elles, mais chacune des approches listées nous paraît pertinente par rapport à notre problématique.

2.1 La recommandation humaine

La technique de filtrage collaboratif équipant les systèmes automatiques de recommendation est conçu pour reproduire le principe du bouche à oreille. Le premier *outil* de recommandation est... l'humain. Des sites comme Digg [14] ou Wikio [15] l'ont bien compris et plutôt que de se fier à l'opinion de quelques éditeurs et leur maigre couverture du Web, ces sites aggrègent les opinions de milliers d'internautes pour décider quelles seront les nouvelles/informations/histoires à mettre en avant sur la page d'accueil.

Au-delà de la déclaration d'amitié et la mise en ligne de tout ou partie de son carnet d'adresses, les services de réseautage deviennent de plus en plus une place où s'échangent des contenus, agrémentés ou non de commentaires. Un service comme Twitter [16], ou encore Facebook [17], est le théâtre de recommandations de pages web, de vidéos, de photos. Dernièrement est apparu un service – et son application smartphone – GetGlue [18], à la fois réseau social et service de recommandation pour des biens cuturels (livres, musiques, émissions télévisuelles, etc.). GetGlue permet aux utilisateurs de choisir des "amis à suivre" ; des notifications indiquent ce que les amis sont en train de découvrir, lire, écouter, goûter (via des "check-in"). Ce sont les amis qui génèrent des recommandations aux autres ; le système permet de choisir ainsi ses propres "sources de recommandations" sur la base de profil public de publications/goûts.

Le réseau d'interconnaissance peut ainsi être *exploité* pour générer de la valeur. L'hypothèse de base étant que le réseau de connaissance véhicule des informations utiles et dignes de confiance. Cette recommandation sociale est un sujet brûlant en Marketing 2.0. Support du bouche à oreille, les services de réseautage deviennent un matériau très convoité pour détecter des leaders d'opinion ou autre personne dite centrale qui propagera une innovation ou un "buzz" rapidement auprès d'un grand nombre de clients potentiels. L'analyse de réseaux sociaux apporte ici son lot d'outils pour détecter les positions clés, mesurer la propagation des nouvelles, le taux de réplication des informations diffusées ou encore détecter des chemins redondants, des liens critiques, qui, s'ils sont supprimés scindent un réseau social en deux sous-réseaux qui ne peuvent plus communiquer...

Les réseaux sociaux et leurs fonctionnalités de recommandation intrinsèques sont dorénavant à prendre en compte dans la conception des services de vente de biens culturels. Par rapport à notre problématique, cependant, l'intérêt du réseautage est indirect. En effet, les utilisateurs qui recevront les recommandations ne sont pas ceux qui sont impliqués dans le réseau

14. http ://www.digg.com
15. http ://www.wikio.fr
16. http ://twitter.com/
17. http ://www.facebook.com/
18. http ://getglue.com/

social étudié ; par contre, exploiter les interactions sur un réseau social (externe) nous permet de bénéficier du choix averti des internautes dans leur sélection de sources ("amis"). Nous verrons que notre méthode est basée sur l'analyse des "relations d'amitié", et donc exploite les rapprochements de profils de goûts que les internautes ont explicitement réalisés eux-mêmes. Rappelons que dans les systèmes de recommandation, le rapprochement de profils est réalisé par un algorithme (calcul de similarité, par exemple corrélation de Pearson).

2.2 Confiance dans les Réseaux Sociaux

Dans la plupart des sites communautaires, la confiance n'est pas directement explicitée par les internautes. Certaines initiatives ont toutefois recueilli explicitement et utilisé les indications de confiance exprimées par les internautes. Jenifer Golbeck est à l'origine du site FilmTrust. Sur ce site de réseautage, les utilisateurs sont invités à explicitement quantifier la confiance qu'ils ont envers leurs amis. L'auteure propose TidalTrust (Golbeck et Hendler, 2006), un algorithme d'inférence de relations de confiance permettant d'attribuer une note de confiance à des pairs qui ne figurent pas dans la liste directe d'amis. Cette confiance entre nœuds (individus) d'un réseau de confiance permet de calculer le poids accordé par un "utilisateur-conseilleur" dans la recommendation d'un film. Le calcul d'inférence de confiance est un calcul local, qui considère le voisinage à distance 2 à partir d'un nœud, ce qui le rend personnalisé et rapide.

Pour prédire la note r_{im} d'un utilisateur i envers un film m, il faut au préalable trouver dans le voisinage étendu de i un ensemble S d'utilisateurs ayant noté déjà ce film m : sont considérés les amis directs, puis les amis des amis, etc. jusqu'à ce que soit trouvé un rayon de voisinage contenant des conseilleurs potentiels. Une fois trouvé ce rayon, tous les conseilleurs s sont considérés et seuls sont ajoutés à l'ensemble S ceux pour qui la confiance inférée t_{is} est supérieure à un seuil ou s'ils font partie des x internautes qui atteignent les meilleurs niveaux de confiance (paramètre x à fixer au départ). Ainsi, la note est-elle calculée de la manière suivante :

$$r_{im} = \frac{\sum_{s \in S} t_{is} r_{sm}}{\sum_{s \in S} t_{is}} \tag{1}$$

Pour les utilisateurs "moyens" qui notent les films de façon similaire à la moyenne, ni le calcul classique du filtrage collaboratif de sélection de voisinage basé sur le coefficient de Pearson, ni le calcul de confiance n'est utile : prédire la note moyenne produit d'aussi bons résultats ! En revanche pour les utilisateurs en désaccord avec la moyenne, le calcul de confiance s'avère plus pertinent pour préserver les particularités d'opinions que le coefficient de corrélation de Pearson qui capture une similarité globale de notations, notations qui sont majoritairement moyennes. En revanche, il est possible de ne pas trouver de conseilleurs, dans les rares cas où aucun des utilisateurs du réseau de confiance atteignables par un utilisateur i n'ait noté un film m. Dans le réseau social FilmTrust, il suffit d'avoir noté un ami pour qu'une recommendation soit possible pour 95% des pairs utilisateur/film.

Dans la même veine, Massa et Avesani (2009) utilisent, pour leur système de recommendation "Trust-aware", un réseau de confiance et une métrique quantifiant à quel point un utilisateur est digne de confiance. Ils comparent deux métriques, l'une locale, le MoleTrust, qui personalise le calcul de confiance pour un internaute donné, d'une façon proche du TidalTrust et l'autre globale, le PageRank, célèbre indicateur de pertinence de Google qui attribue un

indice de "réputation" à chaque nœud d'un réseau, indices globaux et non personalisés (Page et al., 1998). Les tests sur Epinions.com [19] montrent que, sur ce jeu de données, leur algorithme est plus efficace en précision ("accuracy"), i.e. le taux d'erreur est plus faible (Mean Absolute Error) qu'avec le calcul classique de similarités (coefficient de corrélation de Pearson) ; ils montrent également que leur système est moins sensible au problème de démarrage à froid.

Le problème principal de ces approches est de collecter les informations de confiance. A moins d'intégrer la fonctionnalité de notation d'utilisateurs en plus de la notation de films comme dans FilmTrust, les systèmes de recommendation ne disposent pas de jugement de confiance entre internautes. Dans Golbeck (2009), Jenifer Golbeck fait l'hypothèse que dans les réseaux sociaux, les utilisateurs utilisent des caractéristiques de profils pour sélectionner des pairs de confiance. Une corrélation a été établie entre la similarité de profils (de notation de films) et la confiance ; en particulier, si les opinions convergent sur des films notés de manière dite extrême, c'est-à-dire les films dont les notes sont très éloignées de la moyenne, alors la confiance aura tendance à être plus élevée que la moyenne.

Dans notre cas, nous rappelons que les utilisateurs qui recevront les recommandations ne sont pas ceux qui sont impliqués dans le réseau social étudié. Néanmoins, ces approches prennent tout leur sens dans des services tels que GetGlue (cf. sous-section précédente).

2.3 Prédiction de liens

Certains s'intéressent aux techniques de prédiction de liens dans les réseaux sociaux. Potentiellement utilisées à des fins de recommandations, ces techniques prédisent l'évolution de la structure de réseaux. La majorité des approches évaluent la probabilité de l'apparition d'un lien entre deux nœuds en fonction de leur caractéristiques topologiques (approches topologiques), mais d'autres techniques utilisent des caractéristiques des nœuds.

Il s'agit d'une problématique inhérente au domaine scientifique de l'étude de la dynamique des réseaux. Soit $G =< G_1, G_2, ..., G_t >$ un réseau temporel de réseaux sociaux, G_t est le graphe du réseau à l'instant t. La tâche de prédiction de liens consiste à prédire G_{t+1} à partir de G.

David Liben-Nowell et Jon M. Kleinberg Liben-Nowell et Kleinberg (2007) formalisent le problème de la prédiction de liens et développent des approches basées sur des mesures de proximité de nœuds dans un réseau. Ils ont en particulier expérimenté leur méthode sur un réseau de co-publication de papiers scientifiques entre 1994 et 1999 (section physique de e-Print arXiv, www.arxiv.org). Ils montrent qu'un certain nombre de mesures de proximité conduisent à une prédiction de liens qui surpasse nettement le hasard et montrent ainsi que la topologie du réseau contient des informations latentes permettant d'inférer des interactions futures entre les pairs. Les mesures de proximité seront explicitées dans une section ultérieure de ce document.

Sur cette même thématique, Murata et Moriyasu (2008) étudient des sites de questions-réponses qu'ils apparentent à des réseaux sociaux ouverts et dynamiques du web (plus grands en nombre de nœuds que les précédents) et sur lesquels ils testent eux-aussi les mêmes mesures de score basées sur la topologie. Les mesures de proximité utilisant la structure du graphe permettent bien de prédire l'apparition de lien dans de tels réseaux sociaux (même conclusion que les travaux précédents). Ils montrent par ailleurs que les mesures de proximité exploitant à

19. http ://www.epinions.com

la fois la topologie et le poids pré-existants sur les relations (tel que le nombre d'interactions) atteignent de meilleurs résultats que leurs homologues non pondérées.

Comme Huang et al. (2005), Benchettara et al. (2010b) s'intéressent à la recommandation de contenus en modélisant les transactions d'achat ou de notations par des graphes bipartites d'interactions entre utilisateurs et contenus. Les premiers réutilisent les mesures de scores connus des travaux cités dans les paragraphes précédents pour valider la démarche. Les deuxièmes proposent de nouvelles métriques de distance utilisant la nature bipartite du graphe de publications, reliant des auteurs à des publications. Un tel graphe bipartite peut être projeté dans un graphe à plat de deux manières différentes : un graphe d'auteurs où les liens matérialisent la notion de co-écriture et un graphe de publications où les liens indiquent des auteurs communs. Les métriques proposées exploitent la nature bipartite du graphe et considère les deux types de projection. Ils montrent que combiner des métriques directes (calcul de distance faite classiquement sur un graphe à plat) avec des métriques indirectes calculées sur le graphe de la projection duale améliore de façon notable la pertinence de la prédiction de liens de collaboration. Ils testent leurs travaux sur la recommandation de musique sur des données listant des transactions d'achat sur un site de e-commerce et annoncent des gains en précision substantiels (Benchettara et al., 2010a).

Ces travaux exploitent la nature bipartite des graphes (utilisateur - contenus) mais ne tiennent pas compte d'un réseau social sous-jacent (utilisateur - utilisateur), ce qui pour nos travaux est essentiel. D'autre part, notre problématique n'est pas de prédire des relations entre des utilisateurs et des contenus sur le jeu de données analysé, mais d'exploiter des données communautaires avec des utilisateurs externes sur un jeu de données différent. Les utilisateurs sur lesquels l'apprentissage est réalisé ne sont pas les mêmes que ceux à qui les recommandations seront faites.

2.4 Détection de communautés et analyse de comportement

L'analyse de réseaux sociaux inspire depuis peu la recommandation de contenus. Lefeuvre et Cabanac (2010) proposent une méthode combinant des données sociales avec des données thématiques relatives aux articles publiés pour faire de la recommandation de collaboration scientifique. Si l'application est la même que celle décrite dans la section de prédiction de liens (Benchettara et al., 2010b), ici la méthode est différente. L'idée est de définir, en plus d'une fonction thématique calculant une similarité entre chercheurs, trois autres fonctions dites *sociales* : l'inverse du degré de séparation dans un graphe de co-auteurs (le degré de séparation signifie la longueur du plus court chemin), la force de la connectivité valorisant la capacité à joindre une personne via des intermédiaires différents, et enfin, le nombre de conférences communes matérialisant un facteur de rencontre réelle. L'expérimentation auprès de chercheurs montre qu'en introduisant ces fonctions sociales, un système de recommandation de collaborateurs était significativement mieux perçu qu'un moteur uniquement thématique.

Zhao et al. (2010) exploitent les logs d'utilisation "sociale" de contenus (lecture, écriture, commentaire) pour produire une carte sociale. Cette carte présente "qui parle de quoi", elle est issue d'une extraction de communautés. Des groupes contenus sont obtenus selon une méthode de classification, puis ils obtiennent les communautés d'utilisateurs en regroupant les contributeurs associés aux contenus de chaque groupe. La recommandation de contenus et de personnes se fait en interagissant avec la carte : un zoom sur une communauté indique les contenus ordonnés par pertinence, les personnes importantes du groupe, etc.

D'autres travaux vont exploiter les comportements sociaux dans un contexte de recommandation. Lappas et Gunopulos (2010) s'intéressent aux graphes d'annotations de ressources et à la recommandation de contenus. Le Web social contient en effet une grande quantité de données sous forme de texte généré par les utilisateurs. Ces sont des statuts, des commentaires, des tags, etc. Comme nous l'avons évoqué au début de cet article, Poirier et al. (2010) ou encore Bank et Franke (2010), analysent les commentaires textuels que les internautes écrivent à propos de films. La limitation principale de ces approches est la complexité des technologies avancées (analyse textuelles, linguistiques, modèle de connaissances, ontologies...). Les tags (ou étiquettes) sont habituellement choisis de façon informelle et personnelle par les utilisateurs qui annotent une ressource. Carmagnola et al. (2008) et Szomszor et al. (2008) utilisent les tags comme la représentation d'intérêts des internautes et construisent des profils à partir des tags laissés sur plusieurs sites. Certains systèmes de recommandation utilisent les tags sans considération de l'aspect sémantique. La méthode proposée par Zhang et al. (2010) est basée sur le graphe tripartite utilisateur-objet-tag. Les auteurs présentent trois algorithmes pour calculer le niveau d'intérêt d'un utilisateur sur un objet à partir de ce graphe. Les résultats expérimentaux ont démontré que l'algorithme proposé peut donner des recommandations avec une grande précision. Des travaux récents explorent également la détection de communautés sur ces graphes tripartites (Murata, 2009; Suzuki et Wakita, 2009; Murata, 2010).

Toutes ces approches, très différentes les unes des autres, permettent d'analyser des données collectées dans le contexte d'un site et de produire des recommandations aux internautes de ces sites. Nous ne connaissons pas de méthodes basées sur la détection de communautés dans un graphe social externe comme nous le proposons ici.

3 Approche proposée : l'analyse de réseaux sociaux pour la recommandation

Nous avons balayé un certain nombre de travaux exploitant le Web Social dans un contexte de recommandation. Certains de ces travaux s'appuient sur la topologie du graphe d'invidus pour proposer des contenus aux individus de ce même réseau, notamment les méthodes liées aux réseaux de confiance ou la prédiction de liens.

Or l'enjeu de l'étude est ici bien différent. Le but est effet de faire ressortir des recommandations à destination d'utilisateurs "externes", i.e. apprendre à partir des traces laissées par des internautes du web (de Flixster [20] en particulier) des connaissances destinées à être exploitées au sein d'un site/service de VoD tel que Netflix, Orange ou autre.

Nous proposons pour ce faire de produire des associations ou similarités de films basés sur l'usage du site Flixster. L'idée de base de nos travaux est de faire émerger des similaritiés entre films en introduisant une composante sociale dans le calcul. Nous exploitons le fait que les internautes ont choisi d'interagir avec d'autres fans dont les points de vue les intéressent. A la manière du filtrage collaboratif, qui à partir d'usage de consommation/notations de films construit une matrice $user$ x $item$ et conduit à une matrice $item$ x $item$, nous chercherons ici à exploiter le réseau social Flixster pour construire une telle matrice de similarité de films. En considérant un réseau social où les individus se sont inter-connectés ("ajout d'amis"), nous

20. http ://www.flixster.com

supposons pour notre étude que les internautes font eux-mêmes la démarche de trouver des utilisateurs "similaires" dans le jargon de la recommandation.

Nous présentons ici en préliminaires quelques concepts clés qui seront utilisés par notre méthode. Il ne s'agit en aucun cas de présenter de manière exhaustive les travaux d'analyse de réseaux sociaux ou de graphes complexes, qui, ces dernières années, progresssent très rapidement. Nous proposons d'introduire la thématique pour une bonne compréhension de notre méthode, qui, rappelons-le, doit être considérée comme une première approche cherchant à valider nos hypothèses (apprentissage de matrices de similarités de films sur la base d'un réseau social externe).

3.1 Préliminaires sur l'analyse de réseaux sociaux

L'analyse de réseaux sociaux "réels" fait partie du domaine d'étude des " grands graphes réels", des "réseaux complexes" ou encore les "grands graphes terrains". La discipline a pris son essor dans les années 90 avec la généralisation d'Internet, l'analyse de la topologie de l'Internet, la mise en place de réseaux pairs à pairs ou encore l'accès à des données de réseautage. Ces graphes réels présentent des propriétés communes et sont propices à des travaux de recherches communs : Watts et Strogatz (1998), Albert et Barabasi (2002) et Newman (2003) décrivent de manière détaillée ces graphes.

Score, proximité et similarité Quantifier la relation entre nœuds d'un graphe est l'un des problèmes récurrents de l'analyse de graphe. Au delà de la notion de distance (le nombre de liens qui forment le chemin le plus court), il existe d'autres mesures qui permettent d'attribuer un score à un couple de nœuds qu'ils soient directement reliés ou non. Ce type de calcul est clé pour définir la notion de similarité entre nœuds dans un réseau.

Liben-Nowell et Kleinberg (2007) proposent deux types de méthodes (notamment pour la prédiction de liens) : les unes basées sur la notion de voisinage et les autres nécessitant l'ensemble du graphe.

Concernant les méthodes basées sur le voisinage, le principe est basé sur l'hypothèse que deux nœuds x et y ont toutes les chances d'être reliés un jour si leur voisinage $\Gamma(x)$ et $\Gamma(y)$ présentent un fort recouvrement. Si l'entourage est similaire, il y a de fortes chances que ces deux nœuds fassent connaissance, trouvent des affinités (par exemple, co-écrivent un article). Ainsi le $score(x, y)$ peut-il être calculé simplement par le nombre de voisins communs, ou en s'inspirant de la théorie de l'information, grâce au coefficient de Jaccard déterminant le nombre de points communs par rapport à l'ensemble des caractéristiques décrivant les deux individus. Adamic et Adar (2003) utilisent le même principe mais sur la base de profils plus détaillés (incluant du texte ou autres caractérisques décrivant les individus). Leur calcul renforce les points communs z s'ils sont rares pour définir une similarité entre deux individus, ce qui est particulièrement intéressant dans notre contexte de diminuer l'effet populaire de contenus ; c'est ce que nous utiliserons dans notre méthode (cf. équation4 dans une section ultérieure).

Concernant les méthodes basées sur l'ensemble du graphe et non plus locales au voisinage des nœuds, nous trouvons des méthodes classiques en analyse de réseaux sociaux mesurant la proximité ("closeness", "Katz influence"), des méthodes basées sur des algorithmes de parcours aléatoire sur des variations du calcul de centralité ou du PageRank.

Le calcul d'influence de Katz considère l'ensemble des chemins entre chacun des acteurs. Les chemins les plus longs ont un poids plus faible que les plus courts dans la notion d'influence. Un facteur d'atténuation définit comment la longueur affaiblit un chemin. Ainsi pour mesurer un score entre x et y, Liben-Nowell et Kleinberg (2007) proposent la somme de tous les chemins possibles entre x et y en minimisant la contribution les chemins les plus longs. Un score élevé peut correspondre à grande proximité globale, sachant que tous les chemins possibles entre deux acteurs jouent un rôle.

$$score(x,y) = \sum_{\ell=1}^{\infty} \beta^{\ell}.|paths_{x,y}^{\ell}| \tag{2}$$

avec $paths_{x,y}^{\ell}$ étant l'ensemble des chemins de longueur ℓ entre x et y.

Nous pouvons citer également le SimRank (Jeh et Widom, 2002), inspiré du PageRank dans sa conception récursive et sur le principe de renforcer le score d'une paire de nœuds si les voisins de ceux-ci obtiennent eux-mêmes un score élevé deux à deux. D'une complexité en temps de l'ordre de $O(n^2)$, des heuristiques d'élagage consistant à limiter le parcours aléatoire à des nœuds pas trop lointains (dans un voisinage borné) conduit à une complexité quasi linéaire.

$$score(x,y) = C|\Gamma(x)|.|\Gamma(y)| \sum_{a\in\Gamma(x)} \sum_{b\in\Gamma(y)} score(a,b) \tag{3}$$

avec $score(x,x) = 1$ et $C \in \lceil 0,1 \rceil$.

Ce calcul est particulièrement intéressant car il a été testé également sur les graphes bipartites, i.e. les graphes contenant deux types de nœuds (utilisateurs et films par exemple, ou encore utilisateur et tag) et peut potentiellement servir à traiter nos données, dans l'optique où nous souhaiterions tester d'autres méthodes dans le futur (par exemple des méthodes telles que Murata (2009); Suzuki et Wakita (2009); Murata (2010).

Détection de communautés De nombreux états de l'art décrivent la problématique de la détection de communauté : décrire la nature complexe de la structure des réseaux, en dévoiler les zones denses, mettre en évidence la nature hiérarchique de ces zones denses, etc. Ces états de l'art inventorient les différentes techniques explorées (méthodes divisives ou agglomératives, méthodes hiérarchiques, méthode basées sur l'optimisation d'une fonction objective, sur les parcours aléatoires, etc.), ainsi que des méthodes de mesure de qualité de la partition obtenue. Se reporter à Yang et al. (2009), Porter et al. (2009) ou encore Pons (2007).

Un article clé est dû à Newman et Girvan (2004) : leur première méthode de découpage en communautés s'avère être une référence, car elle est intuitive et vient avec une méthode d'évaluation de la qualité de la partition, la modularité. Cette méthode élimine les liens d'intermédiarité forte un à un, et produit une structure hiérarchique de communautés, en partant d'une communauté unique représentant l'ensemble du graphe. Pour choisir la partition la meilleure, la mesure de modularité maximise le nombre de liens intra-communauté et minimise le nombre de liens inter-communautés.

Cependant peu efficace en temps de calcul, des variantes et améliorations ont été proposées ; à l'heure actuelle, cette méthode continue d'inspirer les recherches, et l'on voit apparaître

par exemple des méthodes locales basées sur l'optimisation de la modularité telles que Blondel et al. (2008). D'autres recherches s'orientent plus sur la détection de communautés avec recouvrement (Pizzuti, 2009), partant du principe qu'un nœud puisse appartenir à différentes communautés, d'autres adressent la volumétrie des très grands réseaux (approximation, parallélisation des calculs). Enfin, les graphes valués, orientés, dynamiques ou encore multipartites ne sont pas vraiment exploités à l'heure actuelle.

3.2 Méthode proposée : *Social Popularity*

Soit un jeu de données riche, contenant d'une part un *réseau social*, produit de l'interconnexion d'amis opérée par les membres du réseau, et d'autre part, un *ensemble de films annotés* par ces mêmes membres (notes et commentaires textuels, appelés *reviews* sur le site). Nous proposons une méthode basée sur la popularité dans un contexte de regroupement social dite *Social Popularity*. Cette méthode se découpe en deux étapes :

- Détection de communautés dans le réseau social.
- Calcul de similarité entre films contextuellement aux communautés trouvées.

Détection de communautés dans le réseau social Le but de la première étape est de délimiter des communautés d'internautes qui se sont déclarés des affinités. Nous faisons l'hypothèse ici que naturellement les individus vont rechercher des pairs qui les "intéressent" (concept de la "sélection sociale" évoquée plus haut). Intuitivement, les films que les pairs "sélectionnés" ont annotés auront plus de chance de les interpeler. Ainsi, ce sont les individus eux-mêmes, dans ce contexte, qui définissent leur entourage (et non un algorithme !). Nous faisons également l'hypothèse que l'utilisation seule des contenus des voisins directs est trop limitée. Nous étendons donc la notion de voisinage à celle de communauté.

Nous utilisons ici l'algorithme de Louvain "Fast Unfolding" (Blondel et al., 2008) dont le principe est d'agglomérer de manière ascendante les nœuds en optimisant localement la modularité. La rapidité d'exécution et son adéquatation avec des données de réseautage peu denses ont motivé notre choix. Cette méthode permet de traiter en une complexité quasi linéaire des centaines de millions de nœuds (complexité vérifiée empiriquement sur des graphes peu denses). L'algorithme, agglomératif, débute avec N communautés contenant chacune un des N nœuds du graphe. Pour chaque nœud u, et pour chacun des voisins v de ce nœud, on examine le gain de modularité obtenu en mettant u dans la communauté de v. S'il existe un voisin v pour lequel la modularité augmente, u est placé dans sa communauté (on choisit le voisin v pour lequel ce gain est maximal. La liste des N nœuds est rebalayée tant qu'un nœud peut être déplacé. Lorsque tous les nœuds sont placés, on construit un nouveau graphe à partir des communautés : les nouveaux nœuds représentant les communautés et les liens valués représentant l'ensemble des liens sortants de la communauté vers les nœuds des communautés voisines. Le procédé de placement des nœuds est alors mis en œuvre sur ce nouveau métagraphe. Les étapes s'enchaînent tant que la modularité ne peut être augmentée. Le résultat est uns structure hiérachique de communautés.

Calcul de similarité entre films localement aux communautés trouvées A partir des données d'annotations de films, cette fois, nous produisons, pour chaque utilisateur u, un ensemble des films qu'il a annotés $n \in N_u$. Chaque ensemble de films N_u, relatif à chaque membre d'une

communauté, sera modélisé dans un graphe par une clique (les films d'un ensemble sont tous reliés). En cas de films co-annotés plusieurs fois, le lien entre ces deux films se voit décrit par un poids correspondant au nombre d'utilisateurs ayant noté ces deux films. Nous produisons ainsi un graphe de co-annotation de films par communauté.

A partir de chaque graphe de films co-annotés, nous calculons un score entre chaque paire de films. A la manière de la *Local Popularity* définie par Baluja et al. (2008), nous pourrions définir le score comme étant le poids de l'arc entre deux films voisins. Cette méthode très simple, capture des comportements sociaux reproduits par des moteurs de recommandation comme celui d'Amazon (Les clients qui ont acheté "X" ont également acheté "Y"). Par contre, comme le soulignent les auteurs, les recommandations issues de ce type de métrique seront biaisées par la popularité des films : les premières recommandations seront les films populaires (du moins les plus populaires parmi les films d'une communauté donnée). On peut imaginer que ce ne seront pas les films les plus intéressants à recommander (en partant du principe que les recommandations doivent apporter de la nouveauté par rapport à une liste de popularité classique que l'on trouve sur tous les sites, facile à générer et non attendue dans les recommandations). D'autre part, l'idée est de calculer de manière plus large des *scores* entre des films non directement reliés, tout en minimisant l'effet de popularité des films. Baluja et al. (2008) proposent d'utiliser le coefficient de Jaccard (dans leur algorithme *Local Rare*) déterminant le nombre de voisins communs par rapport à l'ensemble des voisins de chaque film. Nous préférons la mesure d'Adamic/Adar (Adamic et Adar, 2003) basée sur le même principe mais qui renforce d'avantage les points communs s'ils sont rares avec dans notre cas z les voisins communs de deux sommets, et Γx le voisinage du sommet x (cf. équation 4).

$$score(x, y) = \sum_{z \in \Gamma x \cap \Gamma y} \frac{1}{log|\Gamma z|} \tag{4}$$

Au terme de ces deux étapes, nous obtenons, pour chaque communauté issue de l'analyse du réseau social, un ensemble de matrice de similarité entre films.

Nous proposons dans la section suivante une série d'expérimentation exploitant des matrices, et confrontant les similarités de "popularité sociale" avec des calculs de référence, tels que le filtrage collaboratif.

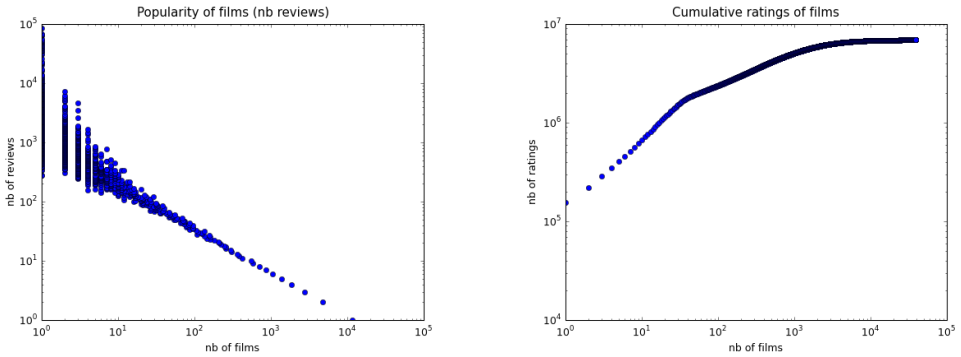
4 Expérimentations

Nous utilisons un jeu de données provenant du site Flixster qui revendique 5 millions d'utilisateurs actifs et 2 millards de commentaires de films (*reviews* en anglais dans le jargon du site). Notre jeu de données a été acquis en juillet 2009, à partir de 100 utilisateurs pris au hasard. A partir de ce noyau, nous avons téléchargé le réseau social à distance 2 (les amis et les amis des amis). En ne gardant que la plus grande composante connexe du graphe social, nous obtenons 459007 utilisateurs et 895794 liens d'amitiés. Le réseau est peu dense ($Densite = 8.5.10^{-6}$) ; il contient environ 10% de liens réciproques, pour un degré maximal de 1108 et un degré moyen de 3.9. La distribution des degrés suit une loi de puissance. Le calcul du coefficient de clustering moyen $C = 0.064$ détermine la proportion de triangles qui existent dans le graphe parmi le nombre possible. Le rayon du graphe est de 7, ce qui signifie qu'il existe au moins un utilisateur qui est proche de tout le monde à une distance inférieure

ou égale à 7. Le calcul de l'*eccentricity* décompte le nombre de nœuds dans ce cas, ici, ils sont au nombre de 97 (on y retrouve en toute logique notre noyau de départ de crawl, ce qui montre bien que la stratégie de collecte impacte fortement la forme du graphe). Ayant calculé exactement les plus courts chemins pour 14380 nœuds, la moyenne des plus courts chemins est 4.520 et le plus court chemin maximal trouvé, i.e. le diamètre est approximativement de 13.0 sur l'échantillon d'un peu plus de 3% de nos nœuds. En utilisant la technique d'approximation de diamètre de Magnien et al. (2009), il suffit d'une centaine de nœuds à fort degré pour arriver à faire converger l'algorithme et estimer le diamètre (même valeur).

Nous sommes conscients que le graphe collecté n'est pas un "réseau social idéal" en ne collectant que les relations à distance 2 des 100 utilisateurs de départ. Dans un contexte de recommandation de films, les études sur les réseaux de confiance nous confortent dans l'idée qu'un voisinage même réduit permet apporter suffisamment d'information pertinente sans introduire trop d'erreurs, nous sommes confiants sur l'intérêt de l'étude.

Nous avons également téléchargé les reviews de ces utilisateurs à propos de 38656 films au total. D'après le nombre de reviews cumulées, nous fixons à 40 le nombre des films les plus *populaires* (cf Fig. 2). Les 40 premiers films cumulent plus d'un million de reviews et ont tous chacun plus de 21248 reviews. Les films dits *mid-populaires* sont entre le rang 41 et 2000 (ils ont plus de 520 reviews chacun). Enfin les films *rares* sont ceux qui occuppent le rang supérieur à 2000. Nous concentrons l'étude présentée ici sur les films rares au nombre de 36656.



a. Le nombre de reviews par film.

b. Cumul du nombre de notes par film.

FIG. 2 – *Le jeu de données collecté sur Flixster présente une Longue Traîne, vue logarithmique (a.). La vue cumulative des notations (b.) par film montre 2 inflexions dans la courbe qui nous permettent de définir trois catégories de popularité de films (popular/mid-popular/rare).*

Nous obtenons 1764784 reviews de films populaires, 4144261 reviews de films mid-populaires et 987160 reviews de films rares, sur lesquelles nous concentrons nos expérimentations décrites ici. Les commentaires rares concernent 62324 utilisateurs ; ceux-ci ont annoté en moyenne 12, 6 films rares (13803 pour le plus prolifique). Le tiers des reviews sont uniques, i.e. les utilisateurs n'ont dans ce cas annoté qu'un film rare (mais ils ont peut-être annotés des films

plus populaires). Une review est constituée d'une note (entre 0 et 5, intervalle 0,5) et d'un commentaire textuel à propos d'un film.

Parmi les utilisateurs de notre réseau social, nous ne gardons que ceux qui ont annoté des films rares. Afin d'évaluer notre méthode nous avons réparti de manière aléatoire les commentaires en deux sous-ensembles d'apprentissage et de test (ratio de 80% / 20%). Parmi les utilisateurs du réseau social, 42541 font partie du jeu de données d'apprentissage.

4.1 Démarche

Le but est d'évaluer, à terme, le pouvoir prédictif de notre méthode dans le contexte d'une recommandation sur un site indépendant (pour d'autres utilisateurs). Dans ce cas, là, les recommandations sont générées en utilisant une matrice de similarité entre films : pour un utilisateur courant, le moteur de recommandations cherche des films non annotés parmi les films similaires des films déjà annotés.

Nous avons généré des matrices de similarité de films avec notre méthode, mais également avec deux autres méthodes de référence (le filtrage collaboratif et la *Local Popularity* légèrement différente de celle mentionnée plus haut, mais dans le même esprit).

$$MAE = \frac{\sum_{i=1}^{N} |predicted_i - real_i|}{N} \tag{5}$$

L'évaluation se fait en mesurant classiquement la *précision* et le *rappel*, ainsi que la *Mean Absolute Error* (MAE) — voir Herlocker et al. (2004) pour une liste exhaustive des mesures d'évaluation. La MAE, équation 5, mesure la déviation absolue moyenne entre une note prédite et une note réelle. Cette mesure est un indicateur de précision de recommantion, c'est-à-dire, dans le cas où il y a une recommandation, jusqu'à quel point est-elle précise. Par contre, la précision et le rappel vont décrire plutôt dans quelle mesure le comportement d'un utilisateur est prédit par le système, si les recommandations faites sont globalement pertinentes, i.e. si elles si retrouvent dans les données réelles.

4.2 Matrices de similarités de films de référence

Filtrage Collaboratif global et rare A partir d'une matrice $user$ x $film$ répertoriant les notes des utilisateurs, nous avons comparé deux à deux les films selon les utilisateurs les ayant notés au moyen du coefficient de corrélation de Pearson. Deux matrices de similarité de films ont été générées en tenant compte d'une part de l'ensemble des reviews du jeu d'apprentissage (populaires, mid-populaires et rares) et d'autre part, en se limitant aux reviews rares de ce même jeu d'apprentissage. Pour chaque cas, pour chaque film, nous gardons les 50 films les plus similaires.

Local Popularity La méthode que nous appellerons *Local Popularity* rejoint dans l'esprit le calcul *Local Rare* décrit par l'équipe de YouTube (Baluja et al., 2008). Il s'agit, à partir d'un graphe bipartite $User$ - $FilmsRares$ de générer un graphe de co-annotations, et d'y calculer des similarités, tout en minimisant l'effet populaire des films les plus annotés. A la place de la distance de Jaccard, nous avons utilisé, comme pour notre méthode, le calcul Adamic/Amar, équation 4.

Ce graphe de co-annotations nécessite un pré-traitement comme le nettoyage des instances de notation unique (pas de co-annotation dans ce cas). Nous générons à partir du jeu d'apprentissage de reviews rares un graphe de co-annotations positives (note supérieure à 2,5) et un graphe de co-annotations négatives.

	nodes	edges	min weight	max weight	nb of scores
Rare positive	7155	457191	5	160	7553308
Rare negative	4758	122769	2	66	7126944

FIG. 3 – *Graphes de co-annotations générés*

Dans ces graphes de co-annotations, nous gardons les liens qui ont un poids supérieur à 5 dans le cas des films positifs, c'est-à-dire les associations de films faites par au moins 5 utilisateurs. Dans le cas des films négatifs, ce seuil a été fixé à 2. Le choix des seuils est fait de manière empirique : les liens gardés ont un poids qui représente 3% du poids maximal (160 dans le cas des liens positifs et 66 dans le cas de liens négatifs).

Le calcul de scores sur ces graphes — selon Admamic/Amar — prend une heure environ pour chaque graphe sur une machine de bureautique [21].

4.3 *Social Popularity*

La première étape de notre méthode consiste à faire une partition du graphe de 459007 nœuds et 895794 arcs. L'algorithme de Louvain produit 85 communautés pour une qualité $Q = 0.6628$ (Modularité initiale : $Q_{init} = -0.0001$, temps de calcul 41 minutes sur notre machine bureautique).

Sur chacune de ces 85 communautés, nous avons généré un graphe de co-annotations positives et un graphe de co-annotations négatives des films rares du jeu d'apprentissage annotés par les membres des communautés. Parmi l'ensemble des individus du réseau social, seuls 42541 se retrouvent dans le jeux de données d'apprentissage et ont annotés un film rare.

Les $85 * 2$ graphes de co-annotations sont ensuite traités à la manière *Local Popularity* décrite plus haut, avec un filtrage des arcs au poids en-deça d'un seuil, puis calcul de scores selon la distance d'Adamic/Adar (équation 4).

L'exploitation de ces matrices de similarité de films contextuelles à des communautés *sociales* d'utilisateurs n'a pour le moment pas été testée de manière approfondie. De manière à simplifier la génération de recommandations, nous avons concaténé l'ensemble des scores en un seule matrice et garder pour chaque films les $k = 250$ plus proches voisins. k a été fixé aléatoirement, et nous pensons par la suite faire varier ce paramètre, d'autant que la valeur choisie nous a fait perdre 40208988 scores, ce qui est non négligeable. Cette phase de filtrage est très longue et a pris 3 semaines sur notre machine bureautique. Au final, ce travail préliminaire nous permet d'obtenir 14399 films rares avec 3278194 scores de similarité calculés sur la base du rapprochement social explicite d'utilisateurs d'une communauté en ligne de fans de cinéma.

21. Machine bureautique équipée d'un Intel(R) Core(TM)2 Duo CPU T9550 @ 2.66GH

4.4 Résultats

Cette section expose nos premiers résultats. Nous avons d'une manière très classique sélectionné les utilisateurs à la fois présents dans le jeu d'apprentissage et le jeu de test. Pour chaque utilisateur (et leur liste de films annotés), nous avons généré k recommandations selon nos 4 méthodes, en mesurant la sensibilité des méthodes sur le nombre de prédictions pris en compte : nous avons fait varier k de 5 à 300.

Pour un utilisateur u, nous cherchons des films candidats i en parcourant la liste des films n qu'il a annotés dans la base d'apprentissage $n \in N$. Nous produisons de manière exhaustive l'ensemble de tous les films similaires aux films de N, en ne gardant que les non présents dans N et en leur attribuant une note selon la méthode de la moyenne pondérée simple (équation 6).

Dans le cas de la *Local Popularity*, les deux matrices de similarité positives et négatives sont distinguées dans les prédictions. Dans le cas où $r_{u,n}$ la note de l'utilisateur courant pour le contenu n est négative, les candidats sont sélectionnés dans la matrice négative (respectivement pour les notes positives).

$$P_{u,i} = \frac{\sum_{n \in N} r_{u,n} w_{i,n}}{\sum_{n \in N} |w_{i,n}|} \tag{6}$$

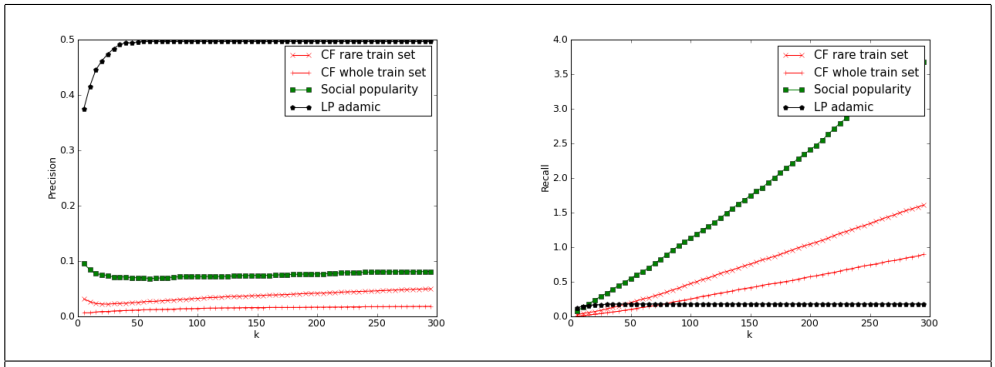


FIG. 4 – *Précision et rappel en fonction du nombre de recommandations émises. La Local Popularity surpasse les autres méthodes mais reste à un niveau de précision très faible : au mieux, à partir de 50 recommandations émises, seules 0.49 % sont présentes dans les listes de films sélectionnées par les utilisateurs. Jusqu'à $k = 300$, le rappel ne cesse d'augmenter, ce qui signifie qu'on trouve des recommandations pertinentes même à un rang élevé dans la liste ordonnée des prédictions. La Social Popularity présente un taux de rappel plus fort que les autres méthodes, tout en restant faible : au mieux 3.67 % des contenus recommandés sont réellement annotés.*

Cette méthode est utilisée pour la prédiction de note utilisant une approche basée sur le contenu (*item-based prediction*) comme dans notre cas. Elle est simple à mettre en œuvre. Elle produit une note dans l'échelle de notation de l'utilisateur u, en utilisant les notes $r_{u,n}$. Le poids $w_{i,n}$ correspond au score de similarité entre les films candidats i et les films n, similarité

calculée précédemment via les techniques de filtrage collaboratif ou la distance d'Adamic/Adar pour les graphes de co-occurrences de films des méthodes *Local* ou *Social Popularity*.

Une fois l'ensemble des films candidats parcourus et associés à une note, une liste ordonnée est produite (selon les notes), et les k premiers films sont proposés en tant que recommandation.

La Fig. 4 montre que la précision du filtrage collaboratif est très faible, ce qui est un problème bien connu, tant le nombre de contenus potentiel est vaste. La *Social Popularity* augmente légèrement cette précision, mais de nombreux contenus recommandés ne se retrouvent pas dans le jeu de test. Par contre, la *Local Popularity* augmente plus largement le nombre de contenus pertinents en diminuant le taux de faux positifs.

Le rappel en revanche est moins bon pour la *Local Popularity*, ce qui revient à dire que le ratio de films recommandés et réellement annotés par rapport au nombre de recommandation est faible. Le filtrage collaboratif est un peu meilleur, notamment si l'apprentissage de similarité s'est restreint aux films rares. Le taux de vrais positifs est un peu meilleur. La *Social Popularity* augmente de façon significative cette probabilité qu'une recommandation se retrouve dans les données réellement annotées par les utilisateurs.

Method	Nb of recommendations
CF whole dataset	200
CF rare dataset	270
Local Popularity	2
Social Popularity	286

FIG. 5 – *Comparatif du nombre de recommandations moyen émises par utilisateur avec* $k = 300$.

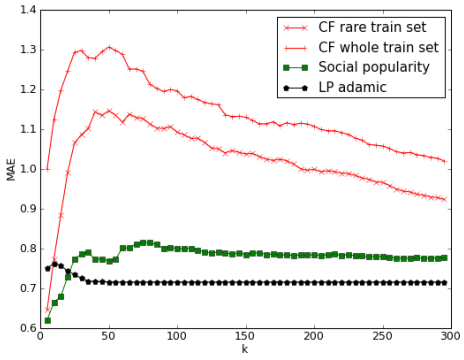


FIG. 6 – *Erreur de prédiction mesurée par la MAE précise la différence de note prédite d'un contenu du jeu de test. Si l'on considère un film sélectionné par l'utilisateur, les méthodes à base de graphe prédisent une note juste avec moins de 0.8 point d'écart, les notes allant de 0 à 5.*

Le Tableau de la Fig. 5 indique que notre méthode émet un nombre de recommandations équivalent au filtrage collaboratif, contrairement au *Local Popularity* qui n'en émet que 2 en moyenne par utilisateur, ce qui est très pauvre.

La Fig. 6 indique de plus que les méthodes locales, utilisant contextes de notations, sont plus précises dans les notes prédites. L'apprentissage sur les films rares uniquement diminue l'erreur pour le filtrage collaboratif ; cette diminution s'accentue lorsqu'en plus les calculs se font contextuellement à des communautés sociales. La méthode *Local Popularity*, légèrement meilleure, a été utilisée un peu différemment en gardant deux matrices de similarités positives et négatives, contrairement aux autres méthodes, ce qui pourraît expliquer la qualité des prédictions : dans un contexte de négation positive, la similarité de deux films peut différer de celle qui est calculée dans un contexte négatif. Ces deux valeurs sont gardées, tandis que pour la méthode *Social Popularity* nous n'avons gardé à ce jour que la meilleure des similarités.

Même si les recommandations générées par notre méthode *Social Popularity* se retrouvent très peu dans les contenus sélectionnés par les utilisateurs (à un niveau comparable aux autres méthodes), par contre elles ont plus de chance d'être justes. La note prédite est elle-même proche de la réalité (écart autour de 0.8 points dans une échelle de 0 à 5), ce qui indique, comme avec la *Local Popularity*, que le score de similarité est plus fiable que celui obtenu avec le coefficient de Pearson.

5 Conclusion

Nous présentons dans cet article une méthode adressant le problème de la recommandation dans la Longue Traîne. Les contenus non populaires, peu annotés, sont difficiles à recommander. Notre originalité est duale et repose sur le fait de capturer la richesse d'annotations du Web Social d'une part, et d'autre part, d'exploiter le fait que les internautes, via un réseau social, sélectionnent eux-mêmes leurs prescripteurs de contenus.

Concernant le premier point, qui répond au problème plus général du démarrage à froid en recommandation, l'idée est de collecter des usages sur le Web Social, de réaliser un apprentissage de connaissances et de réutiliser ces connaissances, ici une matrice de similarité entre films rares, dans d'autres contextes, tels que des services de VoD. Les connaissances apprises sont transposables, grâce à une approche *item-based*.

Le second point, original, revient à analyser un réseau social de fans de cinéma pour faire de la recommandation de films. Nous exploitons les informations données explicitement par les utilisateurs : ce ne sont pas des notes sur les films qui vont servir à établir des similarités entre utilisateurs (filtrage collaboratif), mais le réseau d'amis choisis par affinités de goûts. Nous utilisons une méthode de détection de communautés dans un graphe social, puis nous calculons des similarités de films rares contextuellement à ces communautés — à l'inverse de Park et Tuzhilin (2008) qui regroupent des contenus, et non des individus, pour obtenir un nombre d'annotations suffisant sur les contenus rares.

Ce travail est préliminaire, les premiers résultats indiquent que le réseau social permet de faire des recommandations plus pertinentes (rappel amélioré) en augmentant le nombre de vrais positifs. La précision des recommandations reste faible, mais à un niveau comparable aux autres techniques testées (moins de 0.50% des recommandations se retrouvent parmi les films annotés des utilisateurs). Il conviendrait de confronter de réels utilisateurs au sytème

pour évaluer de manière qualitative les prédictions, en particulier si les films proposés, rares, auxquels ils n'ont pas pensé, ou qu'ils ne connaissent pas tout simplement, sont pertinents.

D'un point de vue technique, nous envisageons de garder la distinction entre similarité positive et négative dans les matrices et tester une méthode *Signed Social Popularity*. Il serait intéressant aussi de tester un nombre k de recommandations plus grand, ainsi que d'autres mesures de similarités dans les graphes de co-annotations de films, d'autres techniques de type prédiction de liens ou détection de communautés dans des graphes tripartites, voire combiner les prédictions issues de différentes méthodes avec un combinateur comme CombMNZ par exemple (Shaw et al., 1994).

Enfin, à terme, il serait intéressant de confronter les matrices de similarité à un jeu de données externes, tels que delui du concours Netflix, de sorte à évaluer l'approche d'externalisation de la phase d'apprentissage.

Remerciements Ce travail a été réalisé dans le cadre d'un contrat avec Orange Labs. Nous remercions notamment l'aide apportée pour collecter les données.

Références

Adamic, L. et E. Adar (2003). Friends and neighbors on the web. *Social Networks 25*(3), 211–230.

Albert, R. et A. Barabasi (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics 74*.

Anderson, C. (2006). *The Long Tail : Why the Future of Business Is Selling Less of More*. Hyperion.

Baluja, S., R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, et M. Aly (2008). Video suggestion and discovery for youtube : taking random walks through the view graph. In *WWW '08 : Proceeding of the 17th international conference on World Wide Web*, New York, NY, USA, pp. 895–904. ACM.

Bank, M. et J. Franke (2010). Social networks as data source for recommendation systems. In W. Aalst, J. Mylopoulos, N. M. Sadeh, M. J. Shaw, C. Szyperski, F. Buccafurri, et G. Semeraro (Eds.), *E-Commerce and Web Technologies*, Volume 61 of *Lecture Notes in Business Information Processing*, pp. 49–60. Springer Berlin Heidelberg. 10.1007/978-3-642-15208-5_5.

Benchettara, N., R. Kanawati, et C. Rouveirol (2010a). Supervised machine learning applied to link prediction in bipartite social networks. *International conference on Advances in Social Networks Analysis and Mining (ASONAM), Odense Danmark*.

Benchettara, N., R. Kanawati, et C. Rouveirol (2010b). A supervised machine learning link prediction approach for academic collaboration recommendation. *International ACM recommender system conference, Barcalona, Spain*, 253–256.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, et E. Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment 2008*(10), P10008.

Candillier, L., K. Jack, F. Fessant, et F. Meyer (2009). State-of-the-art recommender systems. *Collaborative and Social Information Retrieval and Access : Techniques for Improved User Modeling*, 1–22.

Carmagnola, F., F. Cena, L. Console, O. Cortassa, C. Gena, A. Goy, I. Torre, A. Toso, et F. Vernero (2008). Tag-based user modeling for social multi-device adaptive guides. *User Modeling and User-Adapted Interaction 18*, 497–538.

Golbeck, J. (2009). Trust and nuanced profile similarity in online social networks. *ACM Trans. Web 3*(4), 1–33.

Golbeck, J. et J. Hendler (2006). Filmtrust : Movie recommendations using trust in web-based social networks. In *Proceedings of the IEEE Consumer Communications and Networking Conference*.

Herlocker, J. L., J. A. Konstan, L. G. Terveen, et J. Riedl (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst. 22*(1), 5–53.

Huang, Z., X. Li, et H. Chen (2005). Link prediction approach to collaborative filtering. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '05, New York, NY, USA, pp. 141–142. ACM.

Jeh, G. et J. Widom (2002). Simrank : a measure of structural-context similarity. In *KDD '02 : Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 538–543. ACM.

Lappas, T. et D. Gunopulos (2010). Interactive recommendations in social endorsement networks. In *RecSys '10 : Proceedings of the fourth ACM conference on Recommender systems*, New York, NY, USA, pp. 127–134. ACM.

Lefeuvre, A. et G. Cabanac (2010). Confrontation à la perception humaine de mesures de similarité entre membres d'un réseau social académique : enrichissement de la thématique par l'aspect social. In *MARAMI'10 : Actes de la 1ère conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique*.

Liben-Nowell, D. et J. Kleinberg (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology 58*(7), 1019–1031.

Magnien, C., M. Latapy, et M. Habib (2009). Fast computation of empirically tight bounds for the diameter of massive graphs. *J. Exp. Algorithmics 13*, 1.10–1.9.

Massa, P. et P. Avesani (2009). Trust metrics in recommender systems. pp. 259–285.

Murata, T. (2009). Modularities for bipartite networks. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, HT '09, New York, NY, USA, pp. 245–250. ACM.

Murata, T. (2010). Detecting communities from tripartite networks. In *Proceedings of the 19th international conference on World wide web*, WWW '10, New York, NY, USA, pp. 1159–1160. ACM.

Murata, T. et S. Moriyasu (2008). Link prediction based on structural properties of online social networks. *New Generation Computing 26*, 245–257. 10.1007/s00354-008-0043-y.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review 45*, 167–256.

Newman, M. E. J. et M. Girvan (2004). Finding and evaluating community structure in networks. *Phys. Rev. E 69*(2), 026113.

Page, L., S. Brin, R. Motwani, et T. Winograd (1998). The pagerank citation ranking : Bringing order to the web. Technical report, Stanford Digital Library Technologies Project.

Park, Y.-J. et A. Tuzhilin (2008). The long tail of recommender systems and how to leverage it. In *RecSys '08 : Proceedings of the 2008 ACM conference on Recommender systems*, New York, NY, USA, pp. 11–18. ACM.

Piller, F. (2007). Observations on the present and future of mass customization. *International Journal of Flexible Manufacturing Systems 19*, 630–636. 10.1007/s10696-008-9042-z.

Pizzuti, C. (2009). Overlapped community detection in complex networks. In *GECCO '09 : Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, New York, NY, USA, pp. 859–866. ACM.

Poirier, D., F. Fessant, C. Bothorel, M. Boullé, et É. G. D. Neef (2009). Approches statistique et linguistique pour la classification de textes d'opinion portant sur les films. *Revue des nouvelles technologies de l'information (RNTI) 910*, 1 – 202. 9782854289107.
English

Poirier, D., F. Fessant, et I. Tellier (2010). Reducing the Cold-Start Problem in Content Recommendation Through Opinion Classification. In *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference*, Volume 1, Toronto Canada, pp. 204–207.

Pons, P. (2007). *Détection de communautés dans les grands graphes de terrain*. Ph. D. thesis, Université Paris 7 - Denis Diderot.

Porter, M. A., J.-P. Onnela, et P. J. Mucha (2009). Communities in networks. *Notices of the American Mathematical Society 56*(9), 1082–1097, 1164–1166.

Resnick, P. et H. R. Varian (1997). Recommender systems - introduction to the special section. *Commun. ACM 40*(3), 56–58.

Schein, A. I., A. Popescul, L. H. Ungar, et D. M. Pennock (2002). Methods and metrics for cold-start recommendations. In *SIGIR '02 : Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp. 253–260. ACM.

Shaw, J. A., E. A. Fox, J. A. Shaw, et E. A. Fox (1994). Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*, pp. 243–252.

Suzuki, K. et K. Wakita (2009). Extracting multi-facet community structure from bipartite networks. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, Washington, DC, USA, pp. 312–319. IEEE Computer Society.

Szomszor, M., H. Alani, I. Cantador, K. O'Hara, et N. Shadbolt (2008). Semantic modelling of user interests based on cross-folksonomy analysis. In *Proceedings of the 7th International Conference on The Semantic Web*, ISWC '08, Berlin, Heidelberg, pp. 632–648. Springer-Verlag.

Watts, D. et S. Strogatz (1998). Collective dynamics of small-world networks. *Nature 393*, 440–442.

Yang, T., R. Jin, Y. Chi, et S. Zhu (2009). Combining link and content for community detection : a discriminative approach. In *KDD '09 : Proceedings of the 15th ACM SIGKDD international*

conference on Knowledge discovery and data mining, New York, NY, USA, pp. 927–936. ACM.

Zhang, Z.-K., C. Liu, Y.-C. Zhang, et T. Zhou (2010). Solving the cold-start problem in recommender systems with social tags. *CoRR abs/1004.3732*.

Zhao, S., M. X. Zhou, Q. Yuan, X. Zhang, W. Zheng, et R. Fu (2010). Who is talking about what : social map-based recommendation for content-centric social websites. In *RecSys '10 : Proceedings of the fourth ACM conference on Recommender systems*, New York, NY, USA, pp. 143–150. ACM.

Summary

The *Social Popularity* is a new method designed to address the Long Tail recommendation problem, and more generally the cold start problem. Rare items are difficult to recommend because of the lack of usage information such as ratings or reviews. We propose here 1) to capture from the Social Web rich usage information about rare contents, and 2) to exploit the explicit social networking data internet users are offering, giving information on the people they are interested in, their 'movies prescriptors'. Our method first detects communities of users, and then, computes similarity matrices contextually to these communities. The *Social Popularity* method shows encouraging preliminary results: the recall measure is increased, meaning that we can find more true positive predictions than other known techniques. Precision is low, similar to other methods, showing that it is still a challenge to avoid false positive predictions.