Analyse exploratoire d'opinions cinématographiques : co-clustering de corpus textuels communautaires

Damien Poirier, Cécile Bothorel, Marc Boullé

Analyse exploratoire d'opinions cinématographiques : co-clustering de corpus textuels communautaires

Damien Poirier*, Cécile Bothorel*
Marc Boullé*

*TECH / EASY
France Telecom RD
2 avenue Pierre Marzin
22300 Lannion
prénom.nom@orange-ftgroup.com,
http://www.francetelecom.com/fr/groupe/rd/

Résumé. Les sites communautaires sont un endroit privilégié pour s'exprimer et publier des opinions. Le site *www.flixster.com* est un exemple de site participatif sur lequel se rassemblent plus de 20 millions de cinéphiles qui partagent des commentaires sur les films qu'ils ont ou non aimés. Explorer les contenus autoproduits est un challenge pour qui veut comprendre les attentes des internautes. Par une méthode d'apprentissage non supervisée, nous montrerons qu'il est possible de mieux comprendre le vocabulaire utilisé pour décrire des opinions. En particulier, grâce à une méthode de co-clustering, nous montrerons qu'un rapprochement peut être fait entre des films particuliers sur la base de l'usage d'un vocabulaire particulier. L'analyse des résultats peut conduire à retrouver une certaine typologie de films ou encore des rapprochements entre films. Cette étude peut être complémentaire avec des analyses linguistiques des corpus, ou encore être exploitée dans un contexte applicatif de recommandation de contenus multimédias.

1 Introduction

Les avancées technologiques en matière de haut débit favorisent l'apparition de nouveaux services de vente ou location en ligne de fichiers vidéos et musicaux. De tels services se veulent pro-actifs et proposent, en plus des actes promotionnels classiques, des choix personnalisés de films (ou de musique). Des méthodes de recommandation sont déjà utilisées sur certains sites Internet de vente par correspondance (*Amazon*, *Fnac*, *Virgin*, etc.) ou encore sur les plates-formes musicales (*Lastfm*, *Radioblog*, *Pandora*, etc.). Candillier et al. (2007) fait un panorama des techniques de recommandation : qu'elles soient basées sur des notations d'internautes ou des descriptions de contenus (techniques *user-* and *item-based* utilisant le filtrage collaboratif) ou des rapprochements thématiques de profils d'internautes et de descriptions de contenus (filtrage de contenus), voire des techniques hybrides combinant les différentes approches, la problématique reste de gérer les *matrices creuses*. En effet, devant la variété d'un catalogue et le grand nombre d'utilisateurs, le faible nombre de notes qu'un utilisateur donne rend la

comparaison avec d'autres utilisateurs risquée. Parmi les verrous bien connus du domaine de la recommandation figure la difficulté à recueillir des descriptions aussi bien des goûts des utilisateurs (notes, intérêts, usage) que des contenus (métadonnées). Ce problème est pourtant crucial au lancement d'un service (*bootstrapping*), et se repose pour tout nouvel utilisateur, car c'est un challenge d'attirer et fidéliser les clients dès la prise en main du service.

Afin de pallier ces problèmes, une nouvelle voie de recherche est ouverte : l'exploitation de la richesse de l'Internet *ouvert* au service d'un site web *fermé*. Internet est devenu plus que jamais une source de connaissances. Avec le web 2.0 et l'apparition de sites communautaires, les internautes partagent de plus en plus leurs photos, leurs signets, leurs nouvelles, leurs arnaques, leurs opinions. Considérer le web comme un catalogue permet de connaître les goûts d'un nombre considérable d'individus, et nous permet d'envisager à plus ou moins long terme, de décrire des profils type de fans, une typologie des films, ou encore de découvrir de nouveaux descripteurs de films décisifs dans le choix de films.

L'objectif de l'étude exploratoire décrite dans ce papier est de comprendre ce que l'on peut dégager des commentaires de films publiés dans des sites communautaires[1]. Nous chercherons à identifier quel est le vocabulaire révélateur d'opinions, mais également découvrir si un vocabulaire particulier permet de regrouper des films. Nous montrerons que certains mots sont caractéristiques d'un groupe de films précis, tandis que d'autres mots sont beaucoup mieux répartis dans les commentaires de l'ensemble du corpus textuels.

2 Travaux connexes

De nombreux systèmes de mesure d'opinions dans les contenus textuels ont vu le jour récemment. Deux grandes familles de méthodes existent, les méthodes basées sur des techniques de traitement automatique de la langue naturelle (TALN) et celles basées sur l'apprentissage, ces deux types de méthodes pouvant également être combinées.

2.1 Méthodes linguistiques d'analyse d'opinions

Liu et al. (2005) décrivent le système *Opinion Observer* qui analyse finement les commentaires d'usagers dans le but de produire automatiquement des comparatifs de produits commerciaux. Ils ont conçu une méthode de découverte de motifs linguistiques qui permet de trouver le vocabulaire décrivant des critères (comme la qualité des images pour un appareil photo) puis d'en calculer l'orientation. En comptabilisant les scores positifs et négatifs de chaque critère, pour chaque produit, le système produit un rapport détaillé comparant l'opinion générale qui se dégage sur ces produits et facilite ainsi l'achat d'un appareil photo parmi les modèles décrits par les usagers eux-mêmes. *Opinion Observer* est un exemple de système complet basé sur l'identification de mots "porteurs d'opinion" dans les phrases, suivie par un décompte de ces mots. Dave et al. (2003) présentent une méthode plus simple, sans détection de motifs, mais qui, à partir d'un dictionnaire, et selon une échelle d'intensité des mots d'opinions, attribue un score positif ou négatif à chacune des phrases. Grâce aux scores calculés, le système classifie automatiquement des commentaires textuels en 2 classes : positive et négative.

[1]Ce travail entre dans le cadre du projet européen IST Pharos (PHAROS is an Integrated Project co-financed by the European Union under the Information Society Technologies Programme (6th Framework Programme), Strategic Objective "Search Engines for Audiovisual Content" (2.6.3))

Comme beaucoup d'autres systèmes (Morinaga et al., 2002; Turney, 2002; Wilson et al., 2004; Nasukawa et Yi, 2003), ils ont besoin de mots décrivant des opinions. Ce lexique peut être totalement construit à la main ; cependant devant le coût d'un tel procédé, beaucoup de systèmes décrivent des méthodes d'enrichissement plus ou moins automatiques d'une mouture minimale créée manuellement. La constitution de lexique peut se faire grâce à des techniques d'apprentissage. Par exemple, Hatzivassiloglou et McKeown (1997) ou Turney et Littman (2004) utilisent un algorithme non supervisé pour associer de nouveaux mots à des mots pré-sélectionnés. Pereira et al. (1994) et Lin (1998) décrivent des méthodes permettant de découvrir des synonymes en analysant la collocation de mots. Des méthodes linguistiques exploitent l'analyse syntaxique et grammaticale de corpus afin d'étendre le lexique. Citons les travaux de Hatzivassiloglou et McKeown (1997) qui utilisent les conjonctions `and` et `or` pour déduire l'orientation sémantique de vocabulaire ainsi associé à des mots déjà connus. Turney (2002) utilise des motifs un peu plus complexes tels que `adverbe + adjectif non suivis par un nom`. Benamara et al. (2007) se basent sur une classification d'adverbes et attribue des scores à des adjectifs selon la catégorie de l'adverbe auquel ils sont associés.

Citons enfin les travaux de Google (Godbole et al., 2007) ou Hu et Liu (2004) qui utilisent le dictionnaire bien connu WordNet (Miller et al., 1993).

2.2 Méthodes d'apprentissage analysant les opinions

Les systèmes utilisant des méthodes d'apprentissage classifient des commentaires textuels en 2 classes (positive et négative), mais parfois cherchent à prédire des notes de 0 à 5. Ces méthodes de classification supervisées considèrent qu'un commentaire décrit un seul film et cherchent à prédire une note donnée par l'auteur du commentaire.

Beaucoup de méthodes utilisent une préparation linguistique du corpus. Wilson et Wiebe (2003) exposent comment étiqueter les mots porteurs d'opinions par une intensité, après quoi Wilson et al. (2004) testent 3 méthodes d'apprentissage différentes (fréquemment utilisées par les linguistes) : *BoosTexter* (Shapire et Singer, 2000), *Ripper* (Cohen, 1996) et *SVMlight* (la version légère de Support Vector Machine de Joachims (1998)). Cette dernière obtient les meilleurs résultats sur leur corpus annoté. De la même manière, pour caractériser ce qui est ou non apprécié dans chaque phrase, Nigam et Hurst (2004) combinent une technique de *parsing* avec un classifieur bayésien pour associer la polarité à des thématiques.

Pourtant Pang et al. (2002) et Dave et al. (2003) montrent que la préparation des corpus par des lemmatiseurs par exemple, ou encore par la prise en compte des négations, s'avère inutile. Afin de prédire l'opinion de commentaires sur des films, ces deux papiers explorent quelques méthodes d'apprentissage et démontrent qu'elles sont plus performantes que les méthodes de *parsing* suivies d'un décompte comme présentées ci-dessus, avec des résultats de l'ordre de 83% de bonnes prédictions. Les commentaires sont vus comme des sacs de mots. Pang et al. (2002) utilisent un classifieur bayésien naïf et un classifieur maximisant l'entropie.

Dans notre étude exploratoire, nous partons du principe que nous n'avons aucun a priori sur les données. Volontairement, nous n'avons procédé à aucun pré-traitement, et outre le fait que la technique peut de ce fait être utilisée sur toutes les langues, nous prenons l'hypothèse que le vocabulaire dédié aux opinions n'est pas le seul déterminant et utile pour faire des recommandations de films.

3 Technique utilisée

On présente dans cette section une extension au cas non supervisé des modèles en grilles introduits dans le cadre de l'évaluation bivariée pour la classification supervisée (Boullé, 2007c,a).

Après avoir formalisé l'évaluation d'une grille dans le cas de deux variables catégorielles à expliquer, on montre que ce type de grille peut s'interpréter comme un modèle non paramétrique de corrélation entre les valeurs de chaque variable. On décrit ensuite les algorithmes permettant d'optimiser ce type de modèles. On montre enfin qu'il peuvent s'appliquer à l'analyse exploratoire en utilisant un modèle de co-clustering des individus et des variables.

3.1 Groupement de valeurs bivarié non supervisé

On cherche à décrire conjointement les valeurs des deux variables catégorielles à expliquer Y_1 et Y_2, comme illustré sur la figure 1.

D	∅	•	∅	•
C	•	∅	•	∅
B	∅	•	∅	•
A	•	∅	•	∅
	a	b	c	d

{B, D}	∅	•
{A, C}	•	∅
	{a, c}	{b, d}

FIG. 1 – *Exemple de densité jointe pour deux variables catégorielles Y_1 ayant 4 valeurs a, b, c, d et Y_2 ayant 4 valeurs A, B, C, D. Le tableau de contingence sur la gauche ne contient des individus que sur la moitié des cases (marquées •), les autres cases étant vides. Suite au groupement de valeurs bivarié, le tableau de contingence sur la droite permet une description synthétique de la corrélation entre Y_1 et Y_2.*

On introduit en définition 1 une famille de modèles où chaque variable à expliquer est partitionnée en groupes de valeurs. On distribue les individus sur l'ensemble des cellules de la grille bidimensionnelle résultant du produit cartésien des partitions univariées ainsi définies. Cette distribution étant spécifiée, on en déduit par sommation sur les cellules la distribution des individus sur les groupes de valeurs pour chaque variable à expliquer. Il ne reste qu'à spécifier localement à chaque groupe la distribution des individus sur les valeurs du groupe pour obtenir une description complète de la distribution des individus sur les valeurs des deux variables conjointement.

Définition 1. Un modèle de groupement de valeurs bivarié non supervisé est défini par :
- un nombre de groupes pour chaque variable à expliquer,
- la partition de chaque variable à expliquer en groupes de valeurs,
- la distribution des individus sur les cellules de la grille de données ainsi définie,
- la distribution des individus de chaque groupe sur les valeurs du groupe, pour chaque variable à expliquer.

Notations 1.
- N : nombre d'individus de l'échantillon
- V_1, V_2 : nombre de valeurs pour chaque variable (connu)
- J_1, J_2 : nombre de groupes pour chaque variable (inconnu)

- $G = J_1 J_2$: nombre de cellules de la grille du modèle
- $j^{(1)}(v_1), j^{(2)}(v_2)$: index du groupe auquel est rattachée la valeur v_1 (resp. v_2)
- $m_{j_1}^{(1)}, m_{j_2}^{(2)}$: nombre de valeurs du groupe j_1 (resp. j_2)
- $n_{v_1}^{(1)}, n_{v_2}^{(2)}$: nombre d'individus pour la valeur v_1 (resp. v_2)
- $N_{j_1}^{(1)}, N_{j_2}^{(2)}$: nombre d'individus du groupe j_1 (resp. j_2)
- $N_{j_1 j_2}$: nombre d'individus de la cellule (j_1, j_2) de la grille

Un modèle de groupement de valeurs bivarié non supervisé est entièrement caractérisé par le choix des paramètres de partition des valeurs en groupes

$$J_1, J_2, \{j^{(1)}(v_1)\}_{1 \leq v_1 \leq V_1}, \{j^{(2)}(v_2)\}_{1 \leq v_2 \leq V_2},$$

des paramètres de distribution des individus sur les cellules de la grille

$$\{N_{j_1 j_2}\}_{1 \leq j_1 \leq J_1, 1 \leq j_2 \leq J_2},$$

et des paramètres de distribution des individus des groupes sur les valeurs des variables

$$\{n_{v_1}^{(1)}\}_{1 \leq v_1 \leq V_1}, \{n_{v_2}^{(2)}\}_{1 \leq v_2 \leq V_2}.$$

Les nombres de valeurs par groupe sont déduits du choix des partitions des valeurs en groupes, et les effectifs des groupes par comptage des effectifs des cellules de la grille.

Afin de rechercher le meilleur modèle, on applique une approche Bayesienne visant à maximiser la probabilité $P(M|D) = P(M)P(D|M)/P(D)$ du modèle sachant les données. À cet effet, on introduit en définition 2 une distribution a priori sur les paramètres des modèles.

Définition 2. On appelle a priori hiérarchique l'a priori de modèle de densité par grille basé sur les hypothèses suivantes :

- les nombres de groupes de valeurs J_1 (resp. J_2) des variables à expliquer sont indépendants entre eux, et compris entre 1 et V_1 (resp. V_2) de façon équiprobable,
- pour un nombre de groupes donné J_1 de Y_1, toutes les partitions des V_1 valeurs en J_1 groupes sont équiprobables,
- pour un nombre de groupes donné J_2 de Y_2, toutes les partitions des V_2 valeurs en J_2 groupes sont équiprobables,
- pour une grille de taille donnée (J_1, J_2), toutes les distributions multinômiales des N individus sur les G cellules de la grille sont équiprobables,
- pour un groupe donné d'une variable à expliquer donnée, toutes les distributions multinômiales des individus sur les valeurs du groupe sont équiprobables.

Cette distribution a priori sur les paramètres des modèles est hiérarchique, uniforme à chaque étage de la hiérarchie. Pour les distributions multinômiales, les cases vides sont considérées dans la distribution a priori. En utilisant la définition formelle des modèles et leur distribution a priori hiérarchique, la formule de Bayes permet de calculer de manière exacte la probabilité d'un modèle connaissant les données, ce qui conduit au théorème 1.

Théorème 1. *Un modèle d'estimation de densité par grille suivant un a priori hiérarchique est optimal au sens de Bayes si son évaluation par la formule suivante est minimale sur l'ensemble*

de tous les modèles :

$$\begin{aligned}
&\log V_1 + \log V_2 + \log B(V_1, J_1) + \log B(V_2, J_2) \\
&+ \log \binom{N+G-1}{G-1} + \sum_{j_1=1}^{J_1} \log \binom{N_{j_1}^{(1)} + m_{j_1}^{(1)} - 1}{m_{j_1}^{(1)} - 1} + \sum_{j_2=1}^{J_2} \log \binom{N_{j_2}^{(2)} + m_{j_2}^{(2)} - 1}{m_{j_2}^{(2)} - 1} \\
&+ \log N! - \sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} \log N_{j_1 j_2}! \\
&+ \sum_{j_1=1}^{J_1} \log N_{j_1}^{(1)}! + \sum_{j_2=1}^{J_2} \log N_{j_2}^{(2)}! - \sum_{v_1=1}^{V_1} \log n_{v_1}^{(1)}! - \sum_{v_2=1}^{V_2} \log n_{v_2}^{(2)}!
\end{aligned} \tag{1}$$

$B(V, J)$ est le nombre de répartitions de V valeurs explicatives en J groupes (éventuellement vides). Pour $J = V$, $B(V, J)$ correspond au nombre de Bell. Dans le cas général, $B(V, J)$ peut s'écrire comme une somme de nombre de Stirling de deuxième espèce.

La première ligne de la formule (1) regroupe des termes d'a priori correspondant au choix des nombres de groupes J_1 et J_2 et à la spécification de la partition de chaque variable à expliquer en groupes de valeurs. La deuxième ligne représente la spécification de la distribution multinômiale des N individus de l'échantillon sur les G cellules de la grille, suivi de la spécification de la distribution des individus de chaque groupe sur les valeurs du groupe. La troisième ligne représente la vraisemblance de la distribution des individus dans les cellules de la grille, au moyen d'un terme du multinôme. La dernière ligne correspond à la vraisemblance des valeurs localement à chaque groupe pour chacune des variables à expliquer.

3.2 Interprétation

Dans le cas d'une grille comportant une seule cellule, la formule (1) se réduit à :

$$\begin{aligned}
&\log V_1 + \log V_2 + \log \binom{N+V_1-1}{V_1-1} + \log \binom{N+V_2-1}{V_2-1} \\
&+ \log \frac{N!}{n_{v_1}^{(1)}! n_{v_2}^{(1)}! \dots n_{V_1}^{(1)}!} + \log \frac{N!}{n_{v_1}^{(2)}! n_{v_2}^{(2)}! \dots n_{V_2}^{(2)}!}
\end{aligned} \tag{2}$$

ce qui correspond à la probabilité a posteriori d'un modèle multinômial, pour chacune des variables catégorielles Y_1 et Y_2 à expliquer. On peut alors interpréter le modèle de groupement de valeurs bivarié non supervisé de la définition 1 comme un modèle de description de la corrélation entre les deux variables à expliquer. En cas d'indépendance entre les variables, la description des deux variables conjointement se réduit à la somme des descriptions de chaque variable individuellement. Le modèle en grille permet de capturer de façon non paramétrique des corrélations entre les valeurs des variables à expliquer. Le surcoût de description du modèle de corrélation en grille est alors compensé par une description plus concise des valeurs de chaque variable connaissant le modèle de corrélation. Le meilleur compromis est recherché suivant une approche Bayesienne de la sélection de modèles.

Exemple de deux variables catégorielles à expliquer corrélées. Prenons l'exemple de deux variables catégorielles identiques, et d'un modèle en grille M comportant autant de groupes

que de valeurs ($J_1 = V_1$), comme illustré sur la figure 2. Le coût de description de la grille provenant de la formule (1) est alors égal à :

$$2\log V_1 + 2\log B(V_1, V_1) + \log \binom{N + V_1^2 - 1}{V_1^2 - 1} + \log \frac{N!}{n_{v_1}^{(1)}! n_{v_2}^{(1)}! \ldots n_{V_1}^{(1)}!} \qquad (3)$$

d	∅	∅	∅	•
c	∅	∅	•	∅
b	∅	•	∅	∅
a	•	∅	∅	∅
	a	b	c	d

FIG. 2 – *Grille de groupement de valeurs bivarié avec autant de groupes que de valeurs pour deux variables à expliquer identiques* $Y_1 = Y_2$.

En comparant les formules (2) dans le cas d'indépendance et (3) dans le cas d'égalité des variables à expliquer, on observe un surcoût de modélisation dans les termes d'a priori (spécification de chaque groupement de valeurs et spécification de la distribution des individus sur la grille bidimensionnelle). En revanche, le coût de vraisemblance est divisé par deux : les corrélations étant capturées dans le modèle en grille, la description des deux variables à expliquer se réduit à la description d'une seule variable.

3.3 Algorithme d'optimisation

Nous proposons une heuristique gloutonne ascendante d'optimisation, qui, partant d'une solution initiale de partitionnement bivarié aléatoire, procède par fusion itérative des groupes de valeurs tant qu'il y a amélioration du critère. Cet algorithme est précédé d'une étape de pré-optimisation qui consiste à déplacer les valeurs entre les groupes de façon à améliorer le critère. Afin d'améliorer la solution obtenue, une post-optimisation, basée sur le même algorithme de déplacement des valeurs, est également appliquée.

L'heuristique gloutonne a une complexité algorithmique de $O(N^5)$ avec une implémentation naïve. En effet, pour $V \approx N$, il y a $O(N)$ étapes de fusions de groupes effectuée lors de l'heuristique gloutonne, et chaque étape repose sur l'évaluation de $O(N^2)$ fusions de groupes potentielles impliquant des grilles de $O(N^2)$ cellules. On montre dans (Boullé, 2007b) que cet algorithme peut être implémenté avec une complexité algorithmique de $O(N\sqrt{N}\log N)$ en partant d'une solution initiale aléatoire de taille $O(\sqrt{N})$. Pour atteindre cette complexité, on exploite l'additivité du critère d'évaluation des grilles bivariées, qui se décompose sur les caractéristiques de la grille, des variables, des groupes de valeurs et des cellules. On utilise également la nature intrinsèquement creuse des grilles qui comportent au plus N cellules non vides (une par individu) pour une taille de $O(N^2)$ cellules potentielles.

Comme cette heuristique gloutonne est efficace en temps de calcul, nous l'avons incorporée au sein de la méta-heuristique Variable Neighborhood Search (VNS) (Hansen et Mladenovic, 2001), qui consiste essentiellement à appeler l'heuristique principale en partant de solutions aléatoires générées dans le voisinage de la meilleure solution. On obtient ainsi un algorithme de type "anytime", qui permet d'améliorer la solution en fonction du temps de calcul disponible.

3.4 Co-clustering des individus et variables

Un co-clustering (Hartigan, 1972) est défini comme le regroupement simultané des lignes et des colonnes d'une matrice. Dans le cas des jeux de données de faible densité, ayant de nombreux 0 dans le tableau croisé individus x variables, le co-clustering est une technique attractive pour identifier des corrélations entre groupes d'individus et groupes de variables (Bock, 1979; Govaert et Nadif, 2006; Dhillon et al., 2003; Lechevallier et Verde, 2004).

Considérons un jeu de données binaire de faible densité avec N individus, K variables et V valeurs non nulles. Un tel jeu de données peut être représenté sous la forme d'un tableau à V lignes et deux colonnes. Cela correspond a un nouveau *tableau de données* avec deux *variables* nommées "ID Individu" et "ID Variable" où chaque *individu* est un couple de valeurs (ID Individu, ID Variable), comme illustré sur la figure 3.

	V_1	V_2	V_3	V_4	V_5	...
I_1	0	1	0	0	0	...
I_2	0	0	1	1	0	...
I_3	0	1	0	0	0	...
I_4	0	0	0	1	1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

⟶

$ID\ Individu$	$ID\ Variable$
I_1	V_2
I_2	V_3
I_2	V_4
I_3	V_2
I_4	V_4
I_4	V_5
⋮	⋮

FIG. 3 – *Jeu de données binaire de faible densité : depuis la matrice creuse (individus x variables) au tableau bivarié dense.*

En appliquant notre methode de groupement de valeurs bivarié non supervisée à ce tableau bivarié dense, on obtient une grille bivariée basée sur le groupement des individus d'une part, le groupement des variables d'autre part. Le critère d'évaluation du groupement bivarié conduit à maximiser la corrélation entre groupes d'individus et groupes de variables, ce qui correspond à l'objectif général du co-clustering. Il est à noter que les co-clusters sont ici les cellules de la grille, et qu'ils forment une partition sans recouvrement du jeu de données.

4 Résulats sur l'exemple d'analyse d'opinions

Nous avons effectué les premières analyses sur 50 000 commentaires portant indifféremment sur 7 114 films. Le vocabulaire est composé de 27 673 mots différents. En formatant les données sur le modèle de la figure 3, on obtient alors plus de 700 000 *individus* (film, mot).

Les commentaires de films ont subi un pré-traitement minimal, sans lemmatisation, stematisation ni filtrage sur les mots. Les seuls traitements effectués ont été de mettre tous les caractères en minuscule et la ponctuation a été supprimée. Les commentaires traités sont donc tous de la même forme que les exemples suivants :

- *it was really good juss somethin u wouldnt expect from her*
- *my favourite horror film really good story line*
- *omg i love this film soooooooo soppy but sooooooooooo great*

A l'aide des méthodes décrites précedemment, et après une dizaine d'heures de traitements, 162 ensembles de films ont été créés, composés de 3 à 213 films, ainsi que 141 groupes de mots contenant entre 1 et 495 mots. On obtient ainsi environ 20 000 co-clusters (réduction d'un facteur 10 000 de l'espace de données si l'on considère la matrice *Films*Mots*). On produit ainsi un résumé de l'information que nous interprétons dans les paragraphes suivants.

4.1 Clustering de mots

Nous pouvons observer dans les résultats que les mots sont, en général, classés entre termes de même nature. On retrouve par exemple des groupes de mots :

- de liaison :
  - *in film with from he his has by who films most ...*
  - *for all on are out more some can way make also ...*
- porteurs d'opinions :
  - *great story an well amazing brilliant acting cast fantastic excellent done ...*
  - *cool fun awesome awsome entertaining hot flick enjoyable totally line lots ...*
- portant sur la typologie des films :
  - *action thriller packed adventure smart seat exciting ride suspense edge spectacular teriffic ...*
  - *funny hilarious comedy laugh haha laughs funniest laughed jokes funnier laughing ...*
- portant sur la typologie d'acteurs ou personnages :
  - *bond james sean moore connery villain daniel roger spy franchise craig brosnan ...*
  - *johnny depp orlando pirates bloom keira chocolate depps captain pirate knightley willy ...*
- etc.

D'autres caractéristiques ressortent de certains groupes. On retrouve par exemple les mots n'appartenant pas à la langue anglaise classés dans des ensembles communs. On retrouve ainsi des groupes de mots français d'une part et des groupes de mots espagnols d'autre part.
On retrouve aussi des groupes de mots ne contenant qu'un seul terme comme *"it", "a", "of", "to", "movie", etc.*, termes qui n'apportent que peu de sens, voire pas du tout, et qui peuvent être employés quelque soit le contexte dans le langage courant. C'est pourquoi l'outil ne peut les rapprocher d'aucun autre mot.

4.2 Clustering de films

Parmi les résultats du clustering de films, on trouve des groupes séléctionnés selon différentes caractéristiques.

On observe tout d'abord des ensembles où les films sont classés par genre. On retrouve par exemple les catégories de films pour enfants (*The Lion King, Sleeping Beauty, 101 Dalmatians* ...), les "teens-movies" (films d'ado) (*Scary Movie, Big Mommas House, American Pie* ...), les films d'horreur (*Saw, The Grudge, Final destination* ...), etc..

Certains films sont aussi classés en fonction de l'acteur principal ou du réalisateur. C'est le cas par exemple pour un groupe qui ne contient que des films de Johnny Depp : *Pirates of the Caribbean, Charlie and the Chocolate Factory, Edward Scissorhands, Tim Burton Corpse Bride, Sleepy Hollow, Benny & Joon, The Ninth Gate ...*

Enfin on trouve aussi des groupes de films appartenant à la même série comme les différents épisodes de Star Wars ou de James Bond.

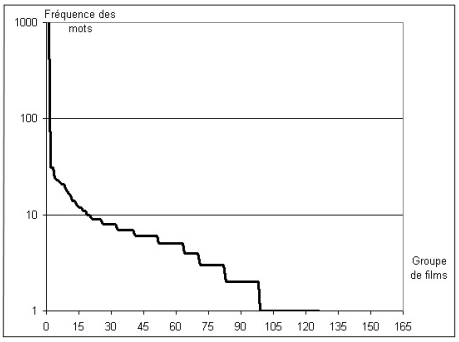

FIG. 4 – *Fréquence des mots du groupe "james bond" par groupe de films.*

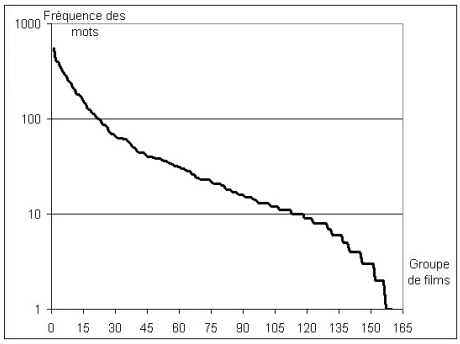

FIG. 5 – *Fréquence des mots du groupe "funny" par groupe de films.*

4.3 Liens entre clusters de mots et clusters de films

Nous pouvons observer au moins trois tendances lorsque que l'on analyse les fréquences de chaque mot d'un même ensemble dans tous les clusters de films. Certain mots, comme ceux référents à un acteur ou personnage, ne sont présents que dans un très faible nombre de clusters de films (exemple avec l'ensemble de mots contenant "*sean, connery, james, bond, etc.*", voir figure 4). D'autres sont présents dans un plus grand nombre, comme les termes décrivant un type de films : comédie, action, policier, etc. (exemple avec l'ensemble de mots contenant "*funny, hilarious, comedy, etc.*", voir figure 5). Et enfin d'autres groupes de mots contiennent des termes présents quasi uniformément dans une grande majorité des clusters de films, c'est le cas pour les ensembles contenant des mots de liaison (*in, film, with, from, he, his, etc.*).

5 Conclusion, perspectives

L'avantage de notre technique de coclustering est qu'elle est fine et fiable, entièrement automatique en ne nécessitant ni paramètre utilisateur, ni connaissance a priori sur le domaine, interprétable et efficace en temps de calcul.

Ces premières analyses de co-clustering de films et de mots nous montrent que les textes produits par les internautes permettent de catégoriser les films selon différents critères. Cette méthode permet notamment de répertorier la plupart des genres existants dans le domaine du cinéma et d'associer ces genres à un vocabulaire précis employé par les amateurs de films. En plus du classement par genres, les films sont catégorisés en fonction des acteurs présents, du réalisateur et de l'appréciation des auteurs des commentaires analysés. Pour chacune de ces caractéristiques, cette méthode nous permet de connaître le vocabulaire s'y rapportant, vocabulaire qui ne paraît pas toujours informatif *a priori* et qu'il serait donc difficile à déterminer par des méthodes linguistiques. Ainsi notre méthode peut être complémentaire avec des analyses linguistiques, en proposant un enrichissement des dictionnaires répertoriant le vocabulaire d'opinion ou de genre cinématographique avec du vocabulaire fiable.

Outre l'exploration même des données, dans un contexte de service de recommandation, la méthode permet naturellement d'associer tout nouveau commentaire (un film qui vient de sor-

tir) à un cluster existant (en minimisant l'impact sur le critère de co-clustering), ce qui permet une analyse automatique des caractéristiques du film (genre, opinion, etc.) par rapprochement à l'existant. On peut également projeter tout nouveau texte (extraits de blogs ou forum) sur le lexique identifié par les clusters de mots et en associer le contenu à un groupe de films.

L'analyse a été faite sur le croisement *Film*Mot* mais on peut également étudier, par exemple, *Utilisateur*Mot* pour identifier une communauté d'utilisateurs qui parlent des films en général avec le meme type de vocabulaire, ou encore *Utilisateur*Film* pour identifier une communauté d'utilisateurs qui commentent les mêmes types de films.

Références

Benamara, F., C. Cesarano, A. Picariello, D. Reforgiato, et V. Subrahmanian (2007). Sentiment analysis : Adjectives and adverbs are better then adjectives alone.

Bock, H. (1979). Simultaneous clustering of objects and variables. In E. Diday (Ed.), *Analyse des Données et Informatique*, pp. 187–203. INRIA.

Boullé, M. (2007a). Optimal bivariate evaluation for supervised learning using data grid models. *Advances in Data Analysis and Classification*. submitted.

Boullé, M. (2007b). Optimization algorithms for bivariate evaluation of data grid models. *Advances in Data Analysis and Classification*. submitted.

Boullé, M. (2007c). Une méthode optimale d'évaluation bivariée pour la classification supervisée. In *Extraction et gestion des connaissances (EGC'2007)*, pp. 461–472.

Candillier, L., F. Meyer, et M. Boullé (2007). Comparing state-of-the-art collaborative filtering systems. International Conference on Machine Learning and Data Mining MLDM 2007, Leipzig/Germany.

Cohen, W. W. (1996). Learning trees and rules with set-valued features.

Dave, K., S. Lawrence, et D. M. Pennock (2003). Mining the peanut gallery : Opinion extraction and semantic classification of product reviews.

Dhillon, I. S., S. Mallela, et D. S. Modha (2003). Information-theoretic co-clustering. In *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003)*, pp. 89–98.

Godbole, N., M. Srinivasaiah, et S. Skiena (2007). Large-scale sentiment analysis for news and blogs. ICWSM'2007 Boulder, Colorado, USA.

Govaert, G. et M. Nadif (2006). Classification d'un tableau de contingence et modèle probabiliste. *Revue des Nouvelles Technologies de l'Information 2*, 457–462.

Hansen, P. et N. Mladenovic (2001). Variable neighborhood search : principles and applications. *European Journal of Operational Research 130*, 449–467.

Hartigan, J. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association 67*(337), 123–129.

Hatzivassiloglou, V. et K. R. McKeown (1997). Predicting the semantic orientation of adjectives.

Hu, M. et B. Liu (2004). Mining and summarizing customer reviews.

Joachims, T. (1998). Making large-scale support vector machine learning practical.

Lechevallier, Y. et R. Verde (2004). Crossed clustering method : An efficient clustering method for web usage mining. Complex Data Analysis, Pékin, Chine.

Lin, D. (1998). Automatic retrieval and clustering of similar words.

Liu, B., M. Hu, et J. Cheng (2005). Opinion observer : Analyzing and comparing opinions on the web.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, et K. Miller (1993). Introduction to wordnet : An on-line lexical database.

Morinaga, S., K. Yamanishi, K. Tateishi, et T. Fukushima (2002). Mining product reputations on the web.

Nasukawa, T. et J. Yi (2003). Sentiment analysis : Capturing favorability using natural language processing.

Nigam, K. et M. Hurst (2004). Towards a robust metric of opinion.

Pang, B., L. Lee, et S. Vaithyanathan (2002). Thumbs up ? sentiment classification using machine learning techniques.

Pereira, F., N. Tishby, et L. Lee (1994). Distributional clustering of english words.

Shapire, R. E. et Y. Singer (2000). Boostexter : A boosting-based system for text categorization.

Turney, P. D. (2002). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews.

Turney, P. D. et M. L. Littman (2004). Unsupervised learning of semantic orientation from a hundred-billion-word corpus.

Wilson, T. et J. Wiebe (2003). Annotating opinions in the world press.

Wilson, T., J. Wiebe, et R. Hwa (2004). Just how mad are you ? finding strong and weak opinion clauses.

Summary

With the Web 2.0 and community sites profusion, Internet users share more and more their opinions. With 20,000,000 users, *www.flixster.com* is an interesting place where web users write reviews about movies and discuss about cinema. Understand this reachable information offers an enormous opportunity for companies eager to better satisfy its customers, especially in the video-on-demand Recommendation context. Non-supervised machine learning techniques are used to explore the huge amount of textual datat. We will show how a new co-clustering algorithm helps in exploring which vocabulary is used to describe opinions. We will explain how such a technique (and the knowledge discovered) may be combined with Natural Language Processing Techniques. We can also identify clusters of movies decribed with specific words, and then propose new similarities between movies, based on the vocabulary used by the fans to describe them, and not classical metadata such as director, date, genre, etc.