Recommandations et prédictions de préférences basées sur la combinaison de données sémantiques et de folksonomie

Pierre-René Lhérisson*,** Fabrice Muhlenbach* et Pierre Maret*

*Univ. Lyon, UJM-Saint-Etienne, CNRS,
Laboratoire Hubert Curien UMR 5516, F-42023 Saint Etienne, France
{pr.lherisson / fabrice.muhlenbach / pierre.maret}@univ-st-etienne.fr
**1D Lab, 5 rue Javelin Pagnon, F-42000 Saint Etienne, France
pierrerene.lherisson@1d-lab.eu

Résumé. Dans les systèmes de recommandation, l'approche du filtrage sur le contenu est revenue en force face à celle du filtrage collaboratif grâce à l'arrivée du paradigme de l'apprentissage profond et des techniques de *word embedding*. Dans cette même veine, l'avènement des folksonomies et du *web* sémantique a apporté une meilleure compréhension des profils des utilisateurs et des caractéristiques des articles à recommander. Dans cet article, nous nous intéressons au domaine musical et nous introduisons un nouveau calcul de mesure de préférence intégrée dans un système de recommandations basées sur le contenu. En testant notre approche sur le jeu de données *Last.fm*, nous montrons que l'utilisation de termes issus d'une folksonomie associés à des informations issues du *web* sémantique permet d'améliorer le processus de recommandation musicale.

1 Introduction

Dans les plates-formes d'e-commerce, il est nécessaire de filtrer la grande masse d'informations disponibles pour proposer automatiquement à chaque consommateur les articles qui sont susceptibles de l'intéresser. Les systèmes de recommandation, qui permettent d'automatiser ce processus (Ricci et al., 2015), procèdent le plus souvent soit suivant le filtrage collaboratif, où les recommandations se font à partir des évaluations déjà effectuées par d'autres utilisateurs sur des articles, soit suivant une approche basée sur le contenu (Ricci et al., 2011) qui utilise les descriptions des articles et des utilisateurs pour créer des profils d'articles et des profils d'utilisateurs.

L'approche de recommandation basée sur le contenu connaît un regain d'intérêt aujourd'hui avec l'arrivée de nouvelles technologies. Néanmoins, pour qu'une approche telle que le *word embedding* soit efficace, il faut que le contenu soit disponible en très grosses quantités sur les articles à recommander. Ceci n'est pas toujours possible suivant les domaines d'application. À titre d'exemple, considérons un système de recommandation associé à un site de *streaming* diffusant de la musique en ligne. Dans cette situation, la masse de contenu textuel exploitable sera importante si les artistes musicaux proposés sont célèbres. Cela ne sera plus vrai dans le cas où le site propose dans son catalogue des artistes musicaux issus de labels indépendants pour lesquels les informations textuelles sont faibles et peu structurées.

Dans cet article, nous nous intéressons au problème de la recommandation musicale dans le cas où une description d'artiste n'est pas disponible pour permettre une application de type *word embedding*. Nous faisons l'hypothèse que les seules informations disponibles sont des annotations d'utilisateurs sous forme de « tags » et qu'à partir de ces informations, associées à des données du *web* sémantique, il est possible d'établir des relations de similarité entre les artistes.

2 Folksonomie et données sémantiques

Lorsque l'on souhaite réaliser une application telle qu'un système de recommandation basée sur le contenu, il faut disposer de données structurées, de qualité, et si possible présentes en grand nombre. Les systèmes basés sur le contenu proposent à un utilisateur des articles similaires à ceux qu'il a aimés dans le passé (Ricci et al., 2011). Cette similitude peut être trouvée par des approches basées sur des heuristiques (avec des techniques empruntées au domaine de la recherche d'information) ou basées sur des modèles (à partir de la construction d'un modèle spécifique pour chaque utilisateur).

La grande force des systèmes collaboratifs est de proposer des évaluations explicites des articles par les différents utilisateurs, et certains systèmes de recommandation basés sur le contenu essaient de retrouver de telles évaluations sur un mode implicite. D'autres approches cherchent à tirer bénéfice des techniques de la fouille de texte et du traitement automatique du langage naturel. Certaines d'entre elles reposent sur les systèmes d'étiquetage (*tagging system*) collaboratif, ou « folksonomie », qui permettent d'enrichir le profil de l'utilisateur. Ces approches ne sont pas toujours efficaces en raison de la liberté laissée aux utilisateurs dans les termes sélectionnables pour annoter les articles (Zhang et al., 2012).

Il existe aussi des approches qui mettent en avant les informations expertes, structurées en graphe, issues des technologies sémantiques, qui ont été adaptées avec succès aux systèmes de recommandation basés sur le contenu (Ricci et al., 2015). La découverte de caractéristiques porteuses de sens dans un texte est fondamentale pour son traitement. Cette connaissance peut être fournie par différentes sources : ontologies, données encyclopédiques non structurées (p. ex. *Wikipédia*) ou sources de données ouvertes liées en ligne (p. ex. *DBpedia*).

Dans ce qui suit, nous montrons comment les systèmes de recommandation sémantiques et basés sur une folksonomie sont des approches qui peuvent s'enrichir mutuellement.

3 Recommandation et préférence

Dans l'approche du filtrage collaboratif, les systèmes de recommandation apportent une réponse à travers un mode de calcul de similarité. Cette similarité peut être donnée de diverses manières. Une première manière (appelée "*item-based collaborative filtering*") consiste à calculer la similarité existant entre les articles à recommander, c'est-à-dire qu'un article similaire à un autre article positivement évalué par l'utilisateur a de fortes chances d'être aussi apprécié. Une deuxième manière (appelée "*user-based collaborative filtering*") exploite les similarités entre comportements d'utilisateurs : un utilisateur dont les évaluations sont similaires à un ou quelques autres utilisateurs aura des chances de suivre la tendance de ce petit groupe et donc il sera possible de trouver la manière dont cet utilisateur appréciera un article non évalué en

étudiant la façon dont son groupe l'a apprécié. La similarité peut être aussi obtenue à partir des évaluations populaires (le suivi de l'opinion générale, positivement ou négativement).

À défaut d'évaluations objectives permettant d'associer un article à un utilisateur donné, il nous faut retrouver un critère tel qu'une mesure de préférence, notée dans la suite *Prf*, avec$Prf_{i,j}$ permettant d'indiquer le niveau de préférence d'un utilisateur i pour un article j.

Dans le domaine musical, nous faisons l'hypothèse que plus un utilisateur préfère écouter un artiste musical donné, plus il aura tendance à écouter des musiques de cet artiste. Cependant, si un artiste musical n'est pas du tout écouté par un utilisateur, une double interprétation de cette valeur nulle d'écoute sera possible : soit il s'agit d'un artiste connu par l'utilisateur et dont celui-ci a volontairement évité d'écouter de la musique parce qu'il ne l'apprécie pas, soit il s'agit d'un artiste non connu de l'utilisateur et, dans ce cas, il va s'avérer pertinent de recommander des musiques de ce dernier, dans la mesure où ces musiques se rapprocheraient de ce qu'aime déjà l'utilisateur-écouteur.

Ce travail propose une manière de déduire la préférence d'un utilisateur pour un artiste dont les musiques n'ont pas été écoutées au moyen d'informations issues de la combinaison de données sémantiques et de folksonomie.

4 Contributions

4.1 Profil d'utilisateur, préférences calculées et préférences prédites

Dans le contexte de la recommandation musicale, nous faisons l'hypothèse que la complémentarité des approches expertes (du *web* sémantique) et des approches non expertes (fournies par une folksonomie) peut s'avérer fructueuse. Dans notre approche, commune à de nombreux systèmes de recommandations basées sur le contenu, nous cherchons à construire un modèle spécifique pour chaque utilisateur afin de pouvoir effectuer les recommandations.

Nous considérons un ensemble d'utilisateurs (les « écouteurs ») $U = \{u_1, \ldots, u_i, \ldots, u_n\}$, un ensemble d'articles (ici, les artistesmusicaux) que nous notons $A = \{a_1, \ldots, a_j, \ldots, a_m\}$, et par $E = \{u_n, a_m\} \in U \times A$ nous indiquons l'ensemble des écoutes effectuées. En se basant sur ces notations, nous définissons par E_i l'ensemble des écoutes d'un utilisateur i donné. Nous cherchons à définir pour chaque écoute $E_{i,j}$ d'un utilisateur i pour un artiste musical j une valeur depréférence *Prf* telle que *Prf* $\in [0, 1]$. À partir des préférences calculées $Prf(i, j)$ pour différents artistes j dont le nombre d'écoutes est non nul, la préférence prédite $\widetilde{Prf}(i, k)$ pour une valeur de préférence inconnue (cas où il n'y a pas d'écoute pour l'artiste musical k concerné) va se calculer de la manière suivante :

$$\widetilde{Prf}(i,k) = \frac{\sum_{j,k \in A_m \cap E_{i,m}} Prf(i,j) \cdot sim(j,k)}{\sum_{j,k \in A_m \cap E_{i,m}} sim(j,k)}$$

4.2 Description des articles à recommander

Pour pouvoir prédire la préférence d'un utilisateur-écouteur i pour un artiste musical k donné, alors que le nombre d'écoutes de cet artiste est nul, il faut pouvoir déduire cette valeur des similarités existant entre cet artiste k et des artistes tels que j où le nombre d'écoutes n'est pas nul, et où la valeur $Prf(i, j)$ a pu être calculée.

Un premier mode de calcul de la similarité entre artistes peut se faire à partir des tags. En faisant l'hypothèse que plus les artistes sont décrits par des tags identiques, plus ils sont considérés comme étant similaires, la similarité entre un artiste j et un artiste k peut facilement être calculée à partir d'une similarité du cosinus. Un second mode de calcul de la similarité entre artistes peut être établi à partir du *web* sémantique. Notre approche consiste à apparier les artistes musicaux de la plate-forme musicale avec les pages correspondantes sur le site de données sémantiques en ligne *DBpedia*. Au sein de ces pages sont récupérées les informations jugées pertinentes pour la description des artistes, telles que le résumé biographique ("*abstract*"), les genres musicaux ("*genre*"), les maisons de disque ("*labels*"), etc.

Ce contenu textuel est traité par l'allocation de Dirichlet latente, ou LDA (Porteous et al., 2008). Ce traitement par LDA permet de retrouver un nombre donné de thèmes latents dans notre corpus (Ponweiser, 2012). À travers les thèmes partagés entre deux artistes j et k, nous pouvons là aussi calculer une mesure de similarité à partir d'une similarité du cosinus.

5 Évaluations expérimentales

5.1 Données expérimentales

Pour évaluer notre approche, nous avons utilisé le jeu de données *Last.fm* de Cantador et al. (2011). Ces données comportent un ensemble initial de $17\ 632$ artistes musicaux, dont nous n'avons conservé que ceux aussi présents sur *DBpedia* et ayant été écoutés au moins une fois par les utilisateurs, soit $m = 8\ 031$ artistes, avec $n = 1\ 753$ utilisateurs, un vocabulaire de $7\ 812$ tags distincts, $72\ 479$ écoutes des artistes conservés et $158\ 444$ annotations faites par les utilisateurs. Les annotations attribuées aux artistes musicaux sont grandement variables et plus un artiste est connu et écouté, plus il est tagué. Ces tags peuvent présenter des éléments très subjectifs de la part des utilisateurs ainsi que des erreurs de saisie. Par exemple, le groupe musical *Morcheeba* a été tagué par un utilisateur donné par : « chillout », « downtempo », « female vovalist » (au lieu de « vocalist »), « electronic », « trip-hop », « dance ».

5.2 Méthodologie

Afin d'étudier les apports respectifs de la folksonomie, des données sémantiques et de la combinaison des deux à la fois, nous avons testé nos approches avec trois modèles :

— **modèle 1** : description des articles par des tags (folksonomie) ;
— **modèle 2** : description sémantique des articles ;
— **modèle 3** : description sémantique des articles + description des articles par les tags.

Chacun des modèles comporte une description des articles (les artistes musicaux) par un mode de calcul de la similarité qui lui est propre, comme indiqué dans la section précédente, et cette description sert de variable d'entrée dans un algorithme d'apprentissage supervisé de type SVM(Cortes et Vapnik, 1995) afin de prédire si l'artiste musical sera écouté ou non par un utilisateur donné. En plus des variables d'entrée qui sont propres à chaque modèle, nous ajoutons d'autres paramètres d'apprentissage tels que la préférence des utilisateurs pour l'artiste musical (préférence calculée ou prédite en cas d'absence d'écoutes), un indice allant de 0 à 1 qualifiant la popularité des artistes musicaux, et une variable binaire correspondant au fait que l'utilisateur a effectivement écouté l'artiste musical en question ou non.

Dans notre protocole d'apprentissage supervisé employé sur les trois modèles, nous sélectionnons aléatoirement 80% des données que nous réservons à l'apprentissage, le test se faisant sur les 20% restants. Les performances obtenues dans la prédiction des valeurs de préférence sont indiquées à travers le calcul de l'erreur quadratique moyenne (RMSE) de $\widetilde{Prf}$ issue de la différence entre la préférence prédite et la préférence effective calculée, du rappel, de la précision et de la F-mesure des trois modèles de recommandation.

6 Résultats et discussions

Le Tableau 1 résume les résultats obtenus par nos calculs de préférences et de recommandations sur les 3 modèles décrits précédemment. Pour analyser ces résultats, il nous semble important de préciser que les résultats présentés sont obtenus à partir de moyennes des modèles individuels de chaque utilisateur. De plus, avec un SVM en validation croisée, l'algorithmen'effectue son apprentissage que sur une seule classe (les seuls artistesécoutés) et cherche à retrouver si un artiste a été écouté ou non par un utilisateur donné. Enfin, la classe des artistes non écoutés (qui n'a pas été apprise) est exagérément sur-représentée par rapport à celle des artistes écoutés (un utilisateur écoute un nombre limité d'artistes du catalogue).

Ce protocole expérimental n'est ainsi pas adapté aux algorithmes d'apprentissage qui discriminent entre deux classes ou plus. Nous avons mis en place des modèles d'apprentissage binaire en utilisant des algorithmes de base (k-PPV, SVM sur 2 classes, forêt aléatoire...) mais les résultats produisent un rappel nul, explicables par le déséquilibre des classes.

	mod. 1 (tag)	mod. 2 (sém.)	mod. 3 (tag + sém.)
RMSE $\widetilde{Prf}$	0.256	0.231	0.232
précision	0.021	0.011	0.023
rappel	0.456	0.544	0.461
F-mesure	0.028	0.018	0.030

TAB. 1 – *Résultats obtenus pour trois modèles de recommandations basées sur le contenu.*

Le Tableau 1 montre que les précisions obtenues dans nos expérimentations sont, dans l'ensemble, relativement faibles, conséquence du déséquilibre des classes en faveur des artistes non écoutés. Le rappel est meilleur pour l'emploi seul des informations de similarité sémantique. Les informations issues de la folksonomie diminuent les résultats en rappel. La folksonomie seule permet difficilement de parvenir à différencier les artistes. En revanche, les tags ont une meilleure correspondance avec les profils des utilisateurs à la manière d'une projection personnelle de l'usager sur l'item « artiste musical ». Les tags ont la propriété de repousser les éléments qui ne sont pas dans la classe des artistes écoutés, mais ils ne permettent pas de retrouver facilement les éléments qui sont dans cette classe.

En conclusion de cette analyse, nous pouvons dire que la combinaison des informations sémantiques et issues de folksonomie semble être un bon compromis car, sans augmenter notablement l'erreur (RMSE), elle améliore la précision sans pour autant trop pénaliser le rappel (le modèle 3 donne la meilleure F-mesure). Ainsi, dans le cas d'artistes non présents sur *DBPedia*, il est pertinent de demander aux utilisateurs de décrire par folksonomie ces artistes.

7 Conclusion et perspectives

Dans cet article, nous avons présenté une méthode permettant d'estimer les préférences des utilisateurs-écouteurs d'une plate-forme de diffusion de musique en *streaming* pour des artistes musicaux. Cette méthode se fonde sur une mesure réalisée à partir des écoutes des utilisateurs. Nous avons présenté des résultats permettant d'étudier le comportement de trois modèles de recommandation réalisés à partir soit de données issues d'une folksonomie, soit de données sémantiques, soit d'une combinaison des deux approches à la fois. Nous avons vu qu'une bonne description des articles à recommander, notamment à travers des informations sémantiques, permet de mieux prédire les prochaines écoutes des utilisateurs.

Ajoutons enfin que notre proposition peut être généralisée à d'autres contextes que le domaine musical. Le travail réalisé ici sur les préférences issues des écoutes peut être transformé afin d'aider à tirer bénéfice du contenu pour la réalisation de systèmes de recommandation dans le cas où, en l'absence de travail collaboratif réalisé par les utilisateurs d'un service en ligne, il n'est pas possible de disposer d'évaluations explicites des différents articles proposés.

Références

Cantador, I., P. Brusilovsky, et T. Kuflik (2011). 2nd workshop on information heterogeneity and fusion in recommender systems (HetRec 2011). In *Proc of RecSys 2011*. ACM.

Cortes, C. et V. Vapnik (1995). Support-vector networks. *Machine Learning 20*(3), 273–297.

Ponweiser, M. (2012). Latent dirichlet allocation in R. Theses / Institute for Statistics and Mathematics 2, WU Vienna University of Economics and Business, Vienna. Diploma Thesis.

Porteous, I., D. Newman, A. Ihler, A. Asuncion, P. Smyth, et M. Welling (2008). Fast collapsed Gibbs sampling for Latent Dirichlet Allocation. In *Proc. of KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA*, pp. 569–577. ACM.

Ricci, F., L. Rokach, et B. Shapira (Eds.) (2015). *Recommender Systems Handbook* (2nd ed.). Springer.

Ricci, F., L. Rokach, B. Shapira, et P. B. Kantor (Eds.) (2011). *Recommender Systems Handbook* (1st ed.). Springer.

Zhang, Z., T. Zhou, et Y. Zhang (2012). Tag-aware recommender systems : A state-of-the-art survey. *CoRR abs/1202.5820*.

Summary

In recommender system, the content-based approach is trending since the arrival of deep learning and word embedding techniques. Otherwise the advent of folksonomies and the semantic web brings a better understanding of user profiles and item features. In this paper, we are focusing on music recommendations and we introduce a new preference index integrated in a content-based recommender system. By testing our approach on *Last.fm* dataset, we show that the use of terms from a folksonomy to describe the music content associated in addition to music information from the semantic weballows to improve the process of music recommendation.