Exploration d'une collection de chansons à partir d'une interface de visualisation basée sur une analyse des paroles

Rémy Kessler, Audrey Laplante, Dominic Forest

Université de Montréal
C.P. 6128, succursale Centre-ville, Montréal H3C 3J7, Canada
{remy.kessler, audrey.laplante, dominic.forest}@umontreal.ca

Résumé. Dans cet article, nous présentons une approche de fouille de textes ainsi qu'une interface de visualisation afin d'explorer une large collection de chansons franaises à partir des paroles. Dans un premier temps, nous collectons paroles et métadonnées de différentes sources sur le Web. Nous utilisons une approche combinant clustering et analyse sémantique latente afin d'identifier différentes thématiques et de déterminer différents descripteurs significatifs. Nous transformons par la suite le modèle afin d'obtenir une visualisation interactive permettant d'explorer la collection de chansons.

1 Introduction

Une grande quantité d'information textuelle sur la musique peut être extraite du Web. On y trouve notamment des données générées par les utitisateurs finaux (p. ex. : tags, critiques), des métadonnées (p. ex. : date de sortie, nom du parolier) et, enfin, les paroles des chansons. Ces informations peuvent souvent être moissonnées au moyen des API qu'offre un nombre croissant de services musicaux sur le Web, de même qu'avec l'aide d'outils développés par la communauté de chercheurs dans le domaine de la recherche d'information musicale. Cependant, les paroles de chansons ont reçu relativement peu d'attention de la part des développeurs des systèmes de repérage pour la musique. Or, la recherche à partir des thèmes abordés dans les paroles peut être pertinente dans certains contextes, pour les chercheurs s'intéressant à la musique populaire ou pour toute personne souhaitant trouver une musique pour un événement particulier (mariage, funérailles). Nous avons donc construit un système d'exploration d'une collection de chansons à partir des paroles. Après la constitution du corpus à partir de données provenant de diverses sources sur le Web, nous avons utilisé des algorithmes de fouille de textes pour en détecter les structures thématiques puis développé une interface de visualisation afin de naviguer dans la collection. Dans cet article, nous expliquons comment les données ont été recueillies et décrivons les différents traitements ayant été appliqués. Nous présentons également l'interface de visualisation qui en résulte.

2 Travaux connexes

Étant donné l'abondance d'informations musicales disponibles en accès libre sur le Web, il n'est pas surprenant de constater qu'un grand nombre de chercheurs ont développé des outils afin de collecter ces informations afin de faciliter le repérage de la musique. Logan et al.

(2004) sont parmi les premiers à exploiter les paroles de chansons dans ce but : ils effectuent une analyse sémantique des paroles afin d'établir le degré de similarité entre artistes. Certains chercheurs utilisent les paroles afin d'effectuer une classification par genre. Ainsi, McKay et al. (2010) utilisent les paroles de chansons en combinaison avec d'autres données (descripteurs acoustiques et symboliques, contenu culturel tiré du Web) pour améliorer la classification automatique par genre. Même si l'approche sous forme de sac de mots reste la plus répandue pour classifier les chansons par genre,Mayer et al. (2008) vont plus loin et utilisent des caractéristiques additionnelles telles que les rimes. Enfin, Kleedorfer et al. (2008) appliquent différentes techniques de fouille ainsi qu'une factorisation en matrices non-négatives (NMF) pour créer des clusters à partir des paroles, avec l'objectif de permettre l'exploration d'une collection de chansons. La visualisation de corpus musicaux a aussi intéressé des chercheurs : plusieurs ont développé des représentations visuelles statiques pour l'exploration de collections où l'espace de représentation est organisé à l'aide des cartes auto-organisatrices de Kohonen. Pour ce faire, la similarité entre chansons est calculée sur la base des fichiers audio (voir *Islands of Music* (Pampalk, 2001) et *MusicRainbow* (Pampalk et Goto, 2006)) ou des paroles dans le cas de *SongWords*, une application pour tablette tactile Baur et al. (2010). Cependant, l'analyse des paroles se limite uniquement à une approche par $Tf*idf$. En conséquence, il semble pertinent de travailler au développement d'interfaces pour l'exploration de collections de chansons à partir des paroles afin de répondre aux besoins décrits précédemment.

3 Constitution d'un corpus de chansons

Avant d'entreprendre la constitution de notre propre corpus, nous avons examiné les jeux de données qui étaient déjà disponibles. Cependant, aucun ne correspondait à nos besoins. Le plus imposant de ces jeux de données est le *Million Song Dataset* (MSD) (Bertin-Mahieux et al., 2011) qui, comme son nom l'indique, contient plus d'un million de chansons. Il est accompagné des paroles de plus de 200 000 chansons sous forme de matrice sac de mots, le *MusiXmatch Dataset*, lequel n'inclut malheureusement qu'une faible portion de chansons en français. Tout d'abord, nous utilisons l'API de musiXmatch pour obtenir une liste d'artistes de pays francophones. Nous collectons par la suite les références pour les albums de ces artistes, ainsi que les titres des chansons pour chaque album. A l'aide des informations obtenues, nous utilisons la librairie jMIR (McKay et al., 2010) ainsi que différentes API [1] pour extraire les paroles. À l'étape subséquente, nous récupérons les métadonnées pour chaque chanson à l'aide de jSongMiner, enrichi pour collecter diverses informations supplémentaires. Comme il s'est avéré difficile de trouver des informations fiables sur la langue des chansons, nous avons dû ajouter une étape de détection de la langue. Nous comparons ainsi les paroles collectées à des antidictionnaires dans plusieurs langues afin de s'assurer de ne récolter que des paroles en langue française. Comme il est courant pour les artistes de sortir plusieurs versions d'une même chanson nous avons dû par ailleurs retirer beaucoup de doublons. Notre corpus contient actuellement les paroles et métadonnées de 4 529 chansons de genres variés et en provenance de divers pays de la francophonie, pour un total de plus d'un million de mots.

1. *LyricWiki*, *Parole.net* et *musiXmatch*

4 Traitement des données

Dans cette section, nous présentons la méthodologie utilisée pour le traitement des données.

4.1 Filtrages et prétraitement linguistiques

Différents prétraitements linguistiques sont effectués sur les données : conversion des majuscules en minuscules, retrait des chiffres et des nombres (numériques et textuels), des accents et des symboles. Afin d'éviter l'introduction de bruit dans le modèle, nous utilisons un anti-dictionnaire classique enrichi de termes extraits de la langue populaire du Québec et de France que l'on retrouve fréquemment dans les chansons (p. ex. : « té » (tu es), « chu » (je suis), « c'te » (cette)). Nous appliquons finalement un processus de lemmatisation simple afin de réduire considérablement les dimensions de l'espace tout en augmentant la fréquence des termes canoniques. Les premiers tests ont cependant montré que, malgré ces traitements, la taille du lexique était toujours importante. Nous présentons donc deux méthodes de sélection : à l'aide de l'analyse sémantique latente (LSA), afin de retenir uniquement les termes les plus représentatifs, et en effectuant une sélection drastique consistant à ne retenir qu'un pourcentage des termes les plus fréquents.

4.2 Clustering des données

Après l'étape de prétraitement, nous appliquons un algorithme de K-moyennes afin d'identifier des clusters de chansons partageant des descripteurs similaires. L'évaluation des résultats d'algorithmes de clustering reste encore aujourd'hui une problématique ouverte importante, particulièrement lorsqu'il n'existe pas de référence. Nous avons évalué le clustering à l'aide de la mesure Silhouette (Rousseeuw, 1987), laquelle permet de mesurer la cohésion ainsi que la distinction des clusters. Nous avons fait varier le nombre de clusters, la mesure de distance (e : Euclidienne, b : Manhattan) ainsi que la méthode de sélection des descripteurs. La baseline est calculée en faisant un tirage aléatoire. Les résultats de l'évaluation (tableau 1) sont assez

Méthode de sélection	Nombre de clusters 2		3		4		5		6		7	
	e	b	e	b	e	b	e	b	e	b	e	b
baseline	-0,58	-0,24	-0,51	-0,69	-0,67	-0,31	-0,75	-0,70	-0,81	-0,49	-0,74	-0,84
1 %	0,00	-0,01	0,08	-0,05	0,12	0,11	0,20	0,03	0,09	0,13	0,10	0,11
2,5 %	0,00	-0,01	0,09	0,03	0,11	0,03	**0,22**	0,19	0,13	0,11	0,12	-0,02
5 %	0,00	0,00	0,08	0,01	0,11	0,09	0,15	0,15	0,11	0,11	-0,02	0,10
LSA	-0,01	0,00	0,05	0,04	0,12	0,13	0,19	0,14	0,05	0,11	0,11	0,03

TAB. 1 – *Évaluation du clustering à l'aide de la mesure Silhouette*

faibles, suggérant des clusters proches les uns des autres et une certaine confusion sur le plan de la classification des chansons. La mesure Silhouette ne permet cependant pas de prendre en compte les chevauchements thématiques ni des spécificités du traitement de données textuelles. La méthode de sélection par LSA présente néanmoins des résultats intéressants étant donné le faible nombre de descripteurs (environ une centaine). Compte tenu des résultats précédents, nous avons choisi pour la suite des expériences de ne retenir que 2,5 % des termes les plus fréquents, fixé à 5 le nombre de clusters et utilisé une distance euclidienne. Une fois l'étape de

clustering terminée, nous utilisons le framework Gensim afin d'indexer chaque cluster séparément et de transformer chaque sous-collection en un modèle LSA (Deerwester et al., 1990). Cette transformation permet de récupérer les mots les plus représentatifs pour chaque cluster, que nous appellerons par la suite les *mots-clés thématiques*, qui peuvent apparaître comme les thèmes associés à chaque cluster. Enfin, nous calculons la distance entre chaque chanson et ces mots-clés, ainsi qu'entre les chansons elles-mêmes avec un seuil minimum (0,4) afin d'éviter d'avoir une visualisation surchargée. Différentes mesures de similarité décrites dans Bernstein et al. (2005) ont été testées : le cosinus, Needleman-Wunsch, Jaro-Winkler. La mesure qui a produit les résultats les plus intéressants était le cosinus.

4.3 Visualisation des données

Plusieurs outils ont été élaborés pour permettre la visualisation des réseaux, tel que Gephi (Bastian et al., 2009) ou Tulip (Auber, 2003). Dans le cadre de ce projet, nous avons utilisé Gephi, un logiciel libre flexible, puissant et particulièrement adapté pour mettre en lumière la structure des associations entre les nœuds d'un réseaux ou d'un graphe. Afin de produire une visualisation des données, nous avons transformé le modèle en un ensemble de nœuds (mots-clés thématiques, artistes et chansons) et de liens tels que :

- Chaque mot-clé thématique est connecté aux chansons du cluster avec un poids déterminé à l'aide du cosinus.
- Chaque chanson est connectée aux autres chansons avec un poids déterminé à l'aide du cosinus (la connexion n'est retenue que si le poids est supérieur au seuil de 0,4).
- Chaque artiste est connecté à ses chansons ainsi qu'aux différents mots-clés thématiques en fonction du nombre d'occurrences de ceux-ci dans les chansons.

Une fois les relations entre tous les éléments définies, une spatialisation est effectuée avec l'algorithme Force Atlas de Gephi. Le résultat est exporté par la suite vers l'interface Web.

4.4 Interface de visualisation

La version actuelle de l'interface contient 4 579 nœuds et 7 789 liens. Les points rouges représentent les mots-clés thématiques. La couleur des nœuds représentant chaque chanson ainsi que celle des liens est déterminée en fonction de son cluster d'appartenance. La sélection d'une chanson ouvre un panneau latéral avec les métadonnées (le nom du chanteur, la couverture de l'album, etc.) ainsi qu'un nuage des termes les plus fréquents de cette chanson. Par ailleurs, les utilisateurs peuvent sélectionner n'importe quel nœud pour obtenir une vue détaillée ou encore zoomer sur une zone particulière. Il est également possible d'effectuer des recherches par mots-clés ou de visualiser chaque cluster séparément. Le système filtre alors la vue courante afin d'afficher le nœud correspondant ainsi que tous les autres nœuds auxquels il est lié. La figure 1 présente une vue générale de la visualisation. Comme mentionné précédemment, la collection a été divisée en 5 clusters. Le tableau 2 présente les mots-clés thématiques pour chacun des clusters : Notamment, on remarque que le premier cluster semble réunir des chansons qui invitent à danser, bouger et chanter. On y trouve entre autres les chansons « Tu vas au bal » et « Père Noël noir », deux chansons à la fois drôles et entraînantes de Renaud. Le deuxième cluster regroupe plutôt des chansons sur l'amour, par exemple la chanson « Mistral gagnant » de Renaud qui parle de l'amour d'un père pour sa fille. En revanche, les mots clés thématiques « revenir » et « regretter » du cluster bleu suggèrent des chansons à propos d'amours déçues ou imaginaires. On y retrouve ainsi la chanson « Me jette pas » de Renaud.

Cluster 1 (orange)	danser, bouger,chanter, ciel, bras
Cluster 2 (rose)	vie, amour, jour, vouloir, aimer
Cluster 3 (bleu)	femme, regretter, revenir, savoir
Cluster 4 (vert)	petit, dire, jamais, homme, nuit
Cluster 5 (cyan)	baby, vivre, blues, bleu, juste

TAB. 2 – *mots-clés thématiques pour les 5 clusters*

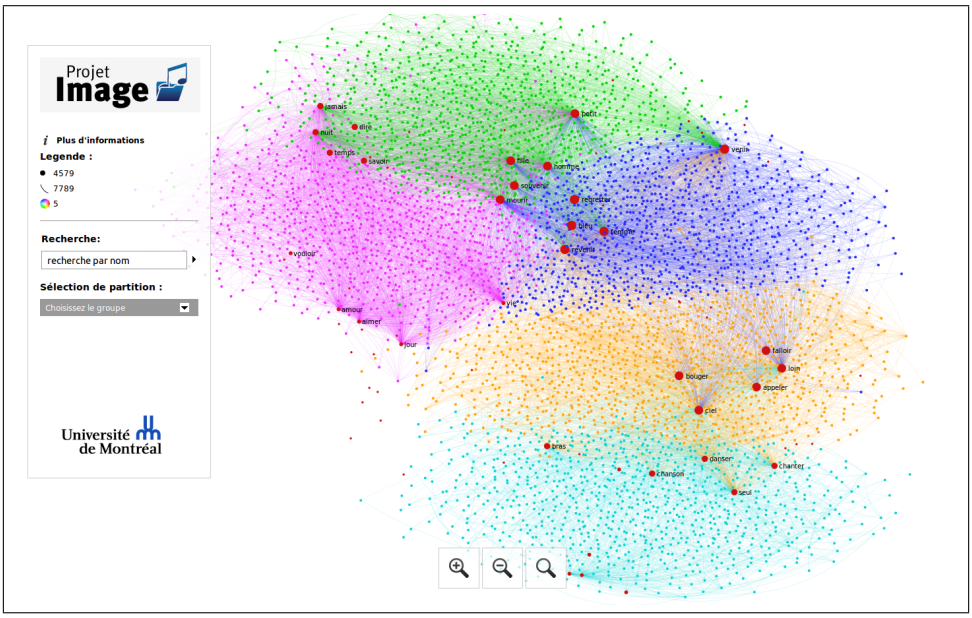

FIG. 1 – *Vue d'ensemble de la collection.*

5 Conclusion et perspectives

Dans cet article, nous avons présenté une méthodologie de fouille de textes permettant d'indexer et de visualiser une collection de chansons. Ce projet est basé sur l'hypothèse que l'analyse, l'organisation et la visualisation des paroles de chansons peuvent permettre aux utilisateurs de naviguer efficacement dans le contenu d'une grande collection musicale. Notre système trouve des applications dans une grande variété de contextes. Nous prévoyons ainsi proposer une visualisation des chansons par décennie et par pays afin de permettre aux chercheurs de comparer les thèmes principaux abordés dans les chansons en fonction de l'époque ou de l'origine de la chanson. En combinant l'analyse thématique des paroles au genre du chanteur, il serait possible aux chercheurs de répondre à une question telle que « Les thèmes abordés dans les chansons par les artistes masculins et féminins diffèrent-ils ? ». Nous prévoyons par ailleurs continuer à augmenter la taille du corpus et évaluer l'ergonomie de l'interface avec des chercheurs s'intéressant à la musique populaire.

Références

Auber, D. (2003). *Tulip : A huge Graph Visualisation Framework*. In Graph Drawing Software, Springer Berlin Heidelberg.

Bastian, M., S. Heymann, et M. Jacomy (2009). Gephi : An Open Source Software for Exploring and Manipulating Networks. In *Int. AAAI conference on weblogs and social media*.

Baur, D., B. Steinmayr, et A. Butz (2010). Songwords : Exploring Music Collections Through Lyrics. In *ISMIR 2010*, pp. 531–536.

Bernstein, A., E. Kaufmann, C. Kiefer, et C. Bürki (2005). SimPack : A generic java library for similarity measures in ontologies. Technical report, University of Zurich.

Bertin-Mahieux, T., D. Ellis, B. Whitman, et P. Lamere (2011). The million Song Dataset. In *ISMIR 2011, Miami, Florida*, pp. 591–596.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, et R. Harshman (1990). Indexing by Latent Semantic Analysis. In *ASIS*, Volume 41, pp. 391–407.

Kleedorfer, F., P. Knees, et T. Pohle (2008). Oh oh whoah ! Towards Automatic Topic Detection in Song Lyrics. In *ISMIR 2008*, pp. 287–292.

Logan, B., A. Kositsky, et P. Moreno (2004). Semantic Analysis of Song Lyrics. In *ICME'04, IEEE International Conference*, pp. 827–830.

Mayer, R., R. Neumayer, et A. Rauber (2008). Rhyme and Style Features for Musical Genre Classification by Song Lyrics. In *ISMIR 2008*, pp. 337–342.

McKay, C., J. A. Burgoyne, J. Hockman, J. Smith, G. Vigliensoni, et I. Fujinaga (2010). Evaluating the Genre Classification Performance of Lyrical Features relative to Audio, Symbolic and Cultural Features. In *ISMIR 2010*, Volume 10, pp. 213–218.

Pampalk, E. (2001). Islands of Music : Analysis, Organization, and Visualization of Music Archives. *Mémoire de master, Vienna University of Technology, Vienna, Austria*.

Pampalk, E. et M. Goto (2006). Musicrainbow : A new User Interface to Discover Artists using Audio-based Similarity and Web-based Labeling. In *ISMIR 2006*, pp. 367–370.

Rousseeuw, P. J. (1987). Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. Volume 20, pp. 53–65.

Summary

In this paper, we present a text mining methodology and an information visualization interface that allows users to browse a large collection of French-language songs based on lyrics. We first harvested lyrics and metadata from various sources on the Web. After data preprocessing, we used clustering and Latent Semantic Analysis to identify a thematic structure and determine significant features. We then transformed the resulting model into a set of nodes and edges to obtain an interactive visualization system for the exploration of our song collection.