Encore des mots, toujours des mots : fouille de textes et visualisation de l'information pour l'exploration et l'analyse d'une collection de chansons en français

Rémy Kessler[1], Dominic Forest[1], Audrey Laplante[1]

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

Abstract

Lyrics, especially lyrics of non-English songs, have rarely been used as an access point for the exploration of a song collection. 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. This system can be used for information retrieval, to find songs on a specific topic, or for research purposes, to identify and visually represent the main topics of a particular collection of songs.

Résumé

Les paroles de chansons, particulièrement en français, ont rarement été utilisées comme point d'entrée afin d'explorer une collection de musique. 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 françaises. Dans un premier temps, nous collectons paroles et métadonnées de différentes sources sur le Web. Après différents prétraitements sur les données, 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 obtenu en un ensemble de nœuds et de liens afin d'obtenir une visualisation interactive permettant d'explorer la collection de chansons. Ce système peut être utilisé pour effectuer des recherches ou afin d'explorer et de visualiser les sujets principaux d'une collection particulière de musiques.

Mots-clés : fouille de textes, interface de visualisation, *clustering*, analyse sémantique latente

1. Introduction

Longtemps négligée car considérée comme un art de seconde classe, la musique populaire a lentement mais sûrement acquis ses lettres de noblesse depuis les années 1950 et est devenue aujourd'hui un sujet d'étude à part entière pour les sociologues, les musicologues et les chercheurs dans les domaines de la communication, des études culturelles et de la culture numérique : ils étudient sa réception, son rôle social, sa signification cultuelle, ses moyens de production et de distribution, etc. (voir p. ex. (Shuker, 2013)). La chanson populaire peut donc être étudiée selon divers aspects, notamment en fonction des thèmes abordés dans leurs paroles. Pensons ainsi aux travaux de Line Grenier, qui s'est intéressée aux thèmes de la citoyenneté et de l'identité dans la chanson québécoise (Grenier, 1997), ou encore aux recherches de Christian Béthune sur l'obscénité et la misogynie dans les paroles des chansons rap (Béthune, 2003). Considérant la grande quantité d'information textuelle sur la musique pouvant être extraite du Web, ce qui inclut métadonnées diverses et paroles de chanson, de

nouvelles avenues pour l'étude de la chanson populaire s'ouvrent aux chercheurs. Au moyen de techniques de fouille de textes, il devient désormais possible de développer des systèmes permettant aux chercheurs d'étendre leur étude des paroles de chansons à de larges corpus. Ces systèmes peuvent en outre être utiles aux usagers finaux. Même si la recherche par sujet n'est certainement pas aussi importante pour la recherche de musique que pour la recherche de documents textuels (Laplante, 2010), celle-ci peut être pertinente dans certains contextes particuliers. On peut supposer, par exemple, qu'un utilisateur qui serait à la recherche de musique pour un événement particulier (un mariage, un enterrement) pourrait être intéressé à explorer une collection musicale à partir des thèmes abordés dans les paroles. De même, les travaux de (Inskip et al., 2010) montrent que les paroles de chansons sont importantes pour les réalisateurs de films et les publicitaires qui se retrouvent souvent à rechercher des chansons dont les paroles seraient liées à la trame narrative du film ou à un spot publicitaire qu'ils produisent. Notre objectif, en nous basant sur les acquis des travaux antérieurs dans ce domaine est double. D'une part, d'un point de vue technique, nous souhaitons développer un système capable de traiter des paroles de chanson en français. D'un point de vue théorique, nous souhaitons, grâce à notre démarche, assister l'analyse d'un corpus de chansons en français, en permettant d'en dégager une structure thématique qui reflète l'évolution des thèmes abordés dans la chanson française. Cet outil permet la recherche et l'exploration d'une large collection de chansons francophones sur la base des thématiques abordées dans leurs paroles. Après une présentation des travaux réalisés par d'autres chercheurs dans ce domaine, nous présentons la façon dont nous avons constitué le corpus. Nous abordons ensuite les différents traitements qui ont été appliqués aux paroles et aux métadonnées du corpus et nous terminons par une présentation de 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. Différentes techniques de fouille de textes sont ensuite employées pour exploiter ces données dans le but de réaliser certaines tâches liées au repérage de la musique, par exemple la classification automatique par genre ou encore la détection d'émotions (voir (Li et al., 2012) pour une revue exhaustive de la littérature sur ce sujet). Plus spécifiquement, les paroles de chansons ont été utilisées pour faciliter le repérage de la musique par quelques chercheurs. (Logan et al., 2004) sont parmi les premiers à exploiter les paroles de chansons dans ce but. Ils effectuent une analyse sémantique des paroles dans le but 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. (Mayer et al., 2008a) adoptent une approche similaire et combinent paroles de chansons et descripteurs acoustiques, aussi pour réaliser la classification par genre. Même si l'approche sous forme de sac de mots reste la plus répandue, (Mayer et al., 2008b) proposent d'utiliser des caractéristiques additionnelles du texte telles que les rimes et les catégories grammaticales, toujours dans le but de classifier les chansons par genre. Des chercheurs ont également démontré la pertinence d'utiliser les paroles pour réaliser d'autres tâches, notamment pour la détection d'émotions (Hu et al., 2009 ; Van Zaanen and Kanters, 2010), parfois en combinaison avec des descripteurs acoustiques (Laurier et al., 2008). (Muller et al., 2007) proposent un système permettant à l'utilisateur de faire une recherche sur les paroles de chansons puis d'accéder directement à la partie de l'enregistrement audio correspondant à sa

requête, ce qui réalisé au moyen d'un alignement automatique entre les paroles et la bande sonore. Quelques outils et méthodes ont également été développés par des chercheurs pour permettre l'extraction de données disponibles sur le Web. (McKay et al., 2010) ont développé lyricFetcher et jLyrics (inclus dans la librarie jMIR[1]) afin de collecter des paroles de chansons et d'en extraire certaines caractéristiques. (Knees et al., 2005) et (Geleijnse et Korst, 2006) ont chacun proposé différentes méthodes visant à identifier le fichier de paroles le plus fidèle parmi un ensemble de doublons. 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 avant de leur assigner manuellement des étiquettes, avec l'objectif de permettre l'exploration d'une collection de chansons. Par ailleurs, de grands efforts ont été déployés pour offrir des interfaces de visualisation aux utilisateurs pour l'exploration de collections musicales. Leur nombre étant assez important, nous ne présentons ici ceux qui nous semblent les plus proches de nos travaux. (Pampalk, 2001) est un pionner dans ce domaine : en 2001, il développe *Islands of Music*, une interface qui prend l'apparence d'une carte géographique où chaque île représente un genre musical. La similarité entre chansons est calculée sur la base des fichiers audio et l'espace de représentation est organisé à l'aide des cartes auto-organisatrices de Kohonen. Plusieurs autres chercheurs ont adopté une approche similaire et développé leur propre interface. (Lübbers, 2005) développe le prototype Sonic SOM, alors que (Pampalk et Goto, 2006) proposent le système *MusicRainbow*. Cependant, bien que les cartes auto-organisatrices puissent être utiles pour créer une représentation visuelle d'une collection que les utilisateurs peuvent facilement comprendre et explorer, cette représentation reste statique. La revue des travaux réalisés dans le domaine indique que les paroles de chansons ont surtout été utilisées pour la classification automatique par genre et pour la détection d'émotions et que très peu d'interfaces ont été développées dans le but de permettre la navigation d'une collection de chansons à partir des thématiques abordées dans les paroles. La seule interface de visualisation utilisant les paroles (par opposition aux fichiers audio) pour estimer la similarité entre les chansons est *SongWords*, une application pour tablette tactile développée par (Baur et al., 2010). Cette application permet d'explorer une collection de chansons à l'aide de cartes auto-organisatrices. Cependant, l'analyse des paroles derrière *SongWords* reste limitée puisqu'elle se base uniquement sur une approche par *Tf*idf*. En conséquence, nous considérons qu'il est nécessaire de poursuivre les efforts dans ce domaine. Notamment, 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 jeu de données SLAC (McKay et al., 2010) inclut paroles de chansons et métadonnées mais ne contient que 250 chansons (dont 90 pièces instrumentales). Le *Million Song Dataset* (MSD) (Bertin-Mahieux et al., 2011) est de loin le plus gros jeu de données de chansons. Comme son nom l'indique, il contient des données sur plus d'un million de chansons, incluant des métadonnées et des descripteurs acoustiques. Bien qu'il y ait un jeu de données complémentaire pour les paroles, le *MusiXmatch Dataset*, contenant les paroles de plus de 200 000 chansons du MSD sous forme de matrice « sac de mots », seule une petite portion des chansons est en français. À l'aide d'une liste de pays francophones, nous avons utilisé les

[1] http://jmir.sourceforge.net/

API de différents entrepôts de données musicales *LyricWiki*, *EchoNest*, *LastFM*, *Paroles.net* et *musiXmatch*, afin de collecter un grand nombre d'informations sur chaque chanson. Le processus de collecte d'informations est détaillé dans (Kessler et al., 2014). Il nous a permis de collecter les paroles de chansons ainsi que de riches métadonnées, incluant les tags sociaux de *Last.fm* et diverses données bibliographiques (p. ex. : nom du parolier et du compositeur, date de sortie). Comme il est courant pour les artistes de sortir un ou plusieurs *singles* avant un album, ainsi que des compilations ou des albums *live*, nous avons dû par ailleurs retirer beaucoup de doublons de notre collection. Le jeu de données final contient les paroles et les métadonnées de 12 109 chansons (après réduction) d'artistes de provenance diverse et couvrant une grande variété de genres musicaux. Le tableau 1 présente quelques statistiques descriptives de notre corpus avec et sans les prétraitements linguistiques décrits en section 4.1.

	Collection
Nombre total de chansons	12 109
sans prétraitement linguistique :	
Nombre total de mots	2 803 194
Nombre total de mots différents	131 727
Moyenne de mots par chanson	227,77
avec prétraitement linguistique :	
Nombre total de mots	1 217 965
Nombre total de mots différents	57 341
Moyenne de mots par chanson	101,28

Tableau 1. Statistiques de la collection

4. Traitement des données

Dans cette section, nous présentons la méthodologie utilisée pour le traitement des données. La figure 3 présente une vue générale de chaque tâche. Les filtrages et les prétraitements linguistiques (étape 1) effectués afin de sélectionner les descripteurs sont décrits dans la section 4.1. Le *clustering* des données (étape 2), l'indexation à l'aide de l'analyse sémantique latente (LSA) (étape 3) ainsi que les mesures de similarité entre les documents sont décrites dans la section 4.2. Les transformations appliquées par la suite ainsi que l'interface de visualisation sont discutées dans les sections 5 et 6.

4.1. Filtrages et prétraitements linguistiques

Nous effectuons dans un premier temps un filtrage sur la collection afin de retirer les chansons en double à l'aide de l'identifiant de *MusicBrainz* ainsi que du titre de chaque chanson. Ce procédé réduit globalement le nombre total de chansons de moitié mais reste cependant nécessaire étant donné qu'il est courant que plusieurs versions d'une même chanson coexistent (voir section 3). Nous avons par ailleurs mis en place une fonction d'évaluation pour ne garder que la version de chaque document contenant le plus de métadonnées. Après cette étape, différents prétraitements linguistiques et processus de

normalisation sont effectués : conversion des majuscules en minuscules, retrait des chiffres et des nombres (numériques et/ou textuels), des accents et des symboles (par exemple « $ »,« # », « * »). Afin d'éviter l'introduction de bruit dans le modèle, nous utilisons un antidictionnaire composé de verbes auxiliaires et d'autres mots fonctionnels couramment utilisés comme marqueurs discursifs ou comme conjonctions de coordination/subordination. Nous avons par ailleurs enrichi l'antidictionnaire de termes extraits de la langue populaire du Québec et de France que l'on retrouve fréquemment dans les chansons (p. ex. : « t'e » (tu es), « chu » (je suis), « te » (cette)). Nous appliquons finalement un processus de lemmatisation simple[2] 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 (plus de 10 000 termes différents). Forest (2009) démontre que les performances des algorithmes de classification sont étroitement liées à la sélection de descripteurs discriminants. Nous présentons donc deux méthodes de sélection : à l'aide de LSA afin de retenir uniquement les termes les plus représentatifs et en effectuant une sélection drastique en ne retenant qu'un pourcentage des termes les plus fréquents. L'ensemble de ces processus nous permet de représenter la collection de chansons selon un modèle en sac de mots (une matrice composée de mots (colonne) pour chaque chanson (ligne)).

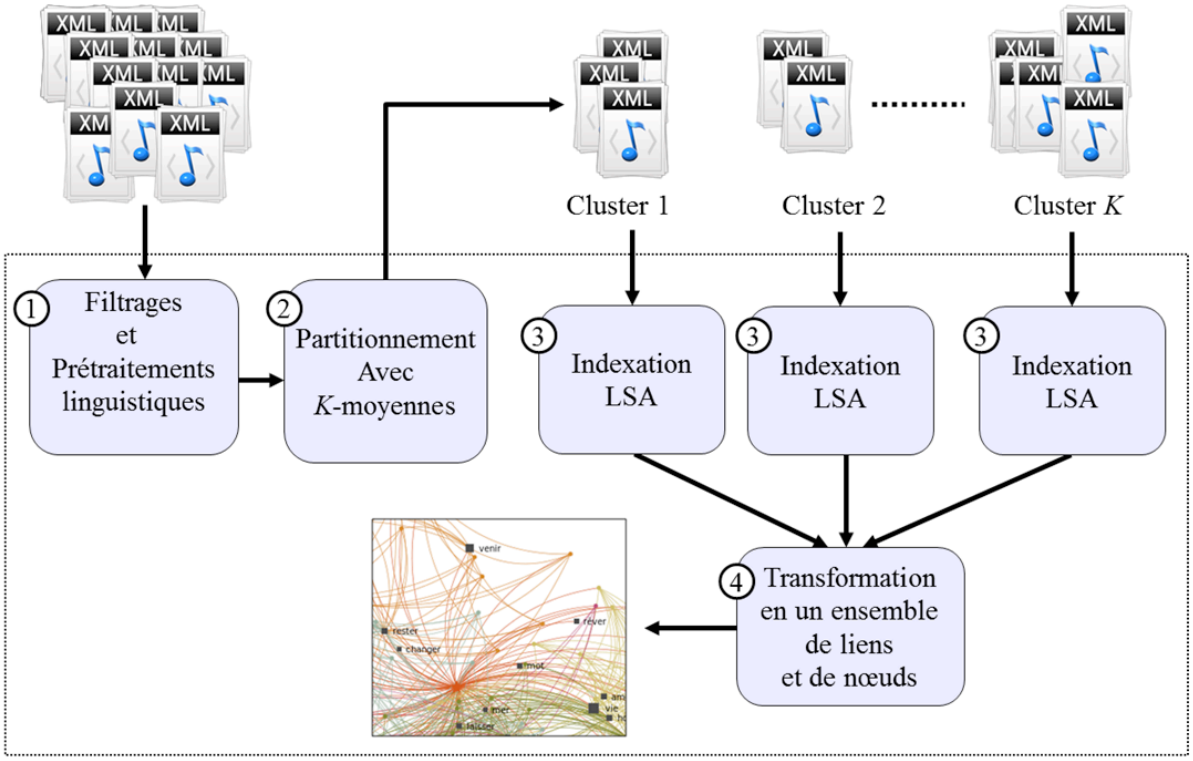

Figure 1. Vue générale des processus de traitement des données

4.2. Clustering des données

Après l'étape de prétraitement, nous appliquons un algorithme de K-moyennes (Hartigan et Wong, 1979) afin d'identifier des *clusters* de chansons partageant des descripteurs similaires. Bien que cette technique soit bien connue dans le domaine du traitement des données

[2] La lemmatisation simple consiste à trouver la racine des verbes fléchis et à ramener les mots pluriels et/ou féminins au masculin singulier.

textuelles, son application demeure un territoire de recherche très fécond. De plus, 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 l'indice de Davies-Bouldin (Davies et Bouldin, 1979). Cette méthode d'évaluation objective ne requiert pas de données de référence. Celui-ci traite chaque classe individuellement et cherche à mesurer à quel point elle est similaire à la classe qui lui est la plus proche. L'indice DB est formulé de la façon suivante :

$$(1) \quad DB = \frac{1}{n} \sum_{i=1}^{n} max \{ \frac{I(c_i) + I(c_j)}{(c_i, c_j)} \}$$

Pour chaque classe i de la partition, on cherche la classe j qui maximise l' « indice de similarité » décrit comme suit :

$$(2) \quad R_{ij} = \frac{(I(c_i) + I(c_j))}{(c_i, c_j)}$$

où $I(c_i)$ représente la moyenne des distances entre les chansons appartenant à la classe $I(c_i)$ et son centre. Et $I(c_i,,c_j)$ représente la distance entre les centres des deux classes c_i et c_j. La meilleure partition est donc celle qui minimise la moyenne de la valeur calculée pour chaque classe. Nous avons fait varier le nombre de *clusters*, ainsi que la méthode de sélection des descripteurs. La base de référence, appelée *baseline* par la suite, est calculée en faisant un tirage aléatoire.

Méthode de sélection	Nombre de *clusters*									
	5	10	15	20	25	30	35	40	45	50
baseline	91,56	156,13	205,46	165,68	228,63	210,51	280,75	299,14	251,41	281,53
1 %	67,47	124,25	155,35	178,56	201,82	225,41	241,35	260,56	278,05	290,35
2,5 %	**43,72**	122,22	154,63	174,33	199,05	216,18	232,45	252,26	272,15	286,69
5 %	58,97	125,74	154,77	175,92	199,52	218,55	235,08	252,46	268,49	285,35
LSA	45,32	142,43	178,88	177,23	189,12	221,63	225,31	286,45	247,17	280,76

Tableau 2. Évaluation du clustering à l'aide de l'indice Davies-Bouldin

Le tableau 2 montre des résultats globalement faibles avec des *clusters* proches les uns des autres et une confusion au niveau de la classification des documents. Nous attribuons cette confusion à trois particularités caractéristiques que nous avons observées dans le corpus : 1) la faible taille des documents (voir tableau 1) rendant difficile la comparaison des documents, 2) la redondance du genre des chansons qui se traduit par une homogénéité du lexique du corpus et 3) la complexité d'extraire des traits discriminants permettant d'identifier des patterns d'information significatifs pour l'utilisateur du système. L'indice Davies-Bouldin ne permet cependant pas de prendre en compte les chevauchements thématiques de certaines chansons ainsi que des spécificités du traitement de données textuelles. La méthode de sélection par LSA présente néanmoins des résultats intéressants compte tenu du faible nombre de descripteurs utilisés (environ une centaine). Étant donné les résultats précédents, nous

avons choisi pour la suite des expériences de ne retenir que 2,5 % des termes les plus fréquents, et de fixer à 5 le nombre de *clusters*. C'est cette combinaison qui nous offre les meilleurs résultats. Une fois l'étape de *clustering* terminée, nous utilisons le *framework* Gensim (Rehurek et Sojka, 2010) afin d'indexer chaque *cluster* séparément et de transformer chaque sous-collection en un modèle LSA (Deerwester et al., 1990). Nous avons choisi d'utiliser LSA car les travaux de (Laurier et al., 2008) suggèrent que LSA produit des résultats de meilleure qualité que les mesures traditionnelles sur des informations textuelles liées à la musique. 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 regroupement. Les observations préliminaires de ces mots-clés thématiques révèlent que les termes qui émergent de cette analyse sont très révélateurs. Afin d'éviter d'avoir une visualisation surchargée, nous avons fixé un seuil minimum (0,4) en dessous duquel le lien n'est pas affiché. Différentes mesures de similarité décrites dans (Bernstein et al., 2005) ont été testées : le cosinus, la distance de recouvrement, Needleman-Wunsch, Jaro-Winkler, la divergence de Jensen-Shannon. Nous avons choisi d'utiliser le cosinus, défini de la façon suivante :

$$(3) \qquad cosinus(j,d) = \frac{\sum_{i=1}^{n} j_i \cdot d_i}{\sqrt{\sum_{i=1}^{n} j_i^2 \cdot \sum_{i=1}^{n} d_i^2}}$$

où j représente les paroles d'une chanson, d'un mot-clé thématique ou les paroles d'une autre chanson, i un terme, j_i et d_i le nombre d'occurrences de i respectivement dans j et d.

5. Visualisation des données

Les techniques de représentation sous forme de réseau ont souvent été utilisées pour représenter visuellement une grande variété de jeux de données, y compris dans le domaine de la musique (voir section 2). Beaucoup d'outils ont été élaborés pour aider à la visualisation des réseaux, tel que Gephi (Bastian et al., 2009), Tulip (Auber, 2003), et Pajek (Batagelj et Mrvar, 1998). Dans le cadre de ce projet, nous avons utilisé Gephi, un logiciel libre flexible, puissant et particulièrement adapté à la visualisation de notre ensemble de données. Afin de visualiser notre ensemble de données, nous avons transformé notre modèle en un ensemble de nœuds et de liens. À l'aide des informations obtenues au cours des traitements précédents, nous définissons trois types de nœuds différents (artistes, mots-clés thématiques et chansons) et plusieurs types de liens :

- Chaque mot-clé thématique est connecté aux chansons du *cluster* avec un poids déterminé à l'aide de la formule (3).
- Chaque chanson est connectée aux autres chansons avec un poids déterminé à l'aide de la formule (3) (la connexion n'est retenue que si le poids est égal ou supérieur au seuil minimal 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 que les relations entre tous les éléments ont été définies, une spatialisation est effectuée en utilisant l'algorithme Force Atlas. Le résultat final est exporté par la suite vers l'interface Web.

6. Interface de visualisation

La version actuelle de l'interface contient 12 499 nœuds et 47 965 liens (voir figure 5). Les carrés gris foncé représentent les mots-clés thématiques tandis que les étoiles écarlates représentent les différents artistes. La couleur des nœuds représentant chaque chanson ainsi que celle des liens partant de ces nœuds est déterminée en fonction du *cluster* d'appartenance de la chanson. La sélection du nœud d'une chanson ouvre un panneau latéral où s'affichent les métadonnées (le nom du chanteur et du compositeur, la couverture de l'album, etc.) ainsi qu'un nuage des termes les plus fréquents de la chanson. Par ailleurs, les utilisateurs peuvent sélectionner n'importe quel nœud pour obtenir une vue plus détaillée (voir figure 4) ou encore zoomer sur une zone particulière de la visualisation afin de voir tous les nœuds. Un menu déroulant permet aux utilisateurs de visualiser chaque *cluster* séparément. Différents boutons permettent par ailleurs de filtrer la visualisation en fonction d'une période particulière (par ex : les années 80, les années 90, etc.). Il est également possible d'effectuer des recherches par mots-clés sur chaque type de nœud. Le système filtre alors la vue courante afin d'afficher uniquement le nœud correspondant ainsi que tous les autres nœuds auxquels il est lié[3]. La figure 4 présente ainsi le sous-réseau pour la chanson intitulée « Petite fille du soleil » du chanteur Christophe. En plus du nœud de la chanson, trois mots-clés thématiques liés à cette chanson sont affichés : « petit », « vie » et « jour ».

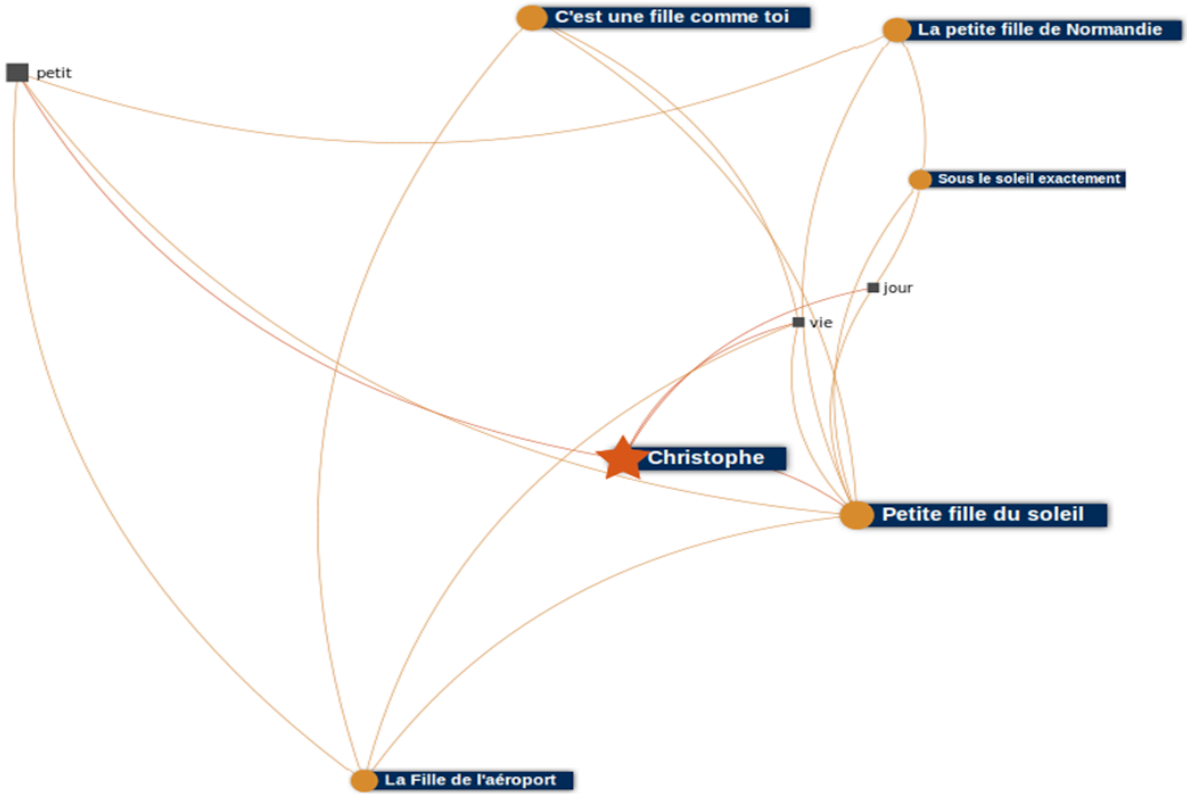

Figure 2. Sous-réseau de la chanson « Petite fille du soleil ».

Ces mots apparaissent tous dans les paroles de cette chanson et sont assez fréquents pour être inclus dans le nuage. Bien que ce sous-réseau soit proche du mot-clé thématique « amour » dans la visualisation d'ensemble, ce terme n'est pas affiché lorsque nous choisissons cette chanson particulière car elle n'y est pas directement reliée (le thème de l'amour est bien

[3] Une démonstration interactive du système sera présentée lors de la communication.

présent dans la chanson mais le mot « amour » n'est pas utilisé). Nous notons également que cette chanson est étroitement liée à trois autres chansons : « La petite fille de Normandie » de Claude Barzotti, « C'est une fille comme toi » de Johnny Halliday et « La fille de l'aéroport » de Patrick Bruel. La visualisation montre aussi une connexion plus faible avec une quatrième chanson, « Sous le soleil exactement » de Serge Gainsbourg. En observant les paroles complètes de notre chanson sélectionnée ainsi que celles des trois chansons les plus proches, on remarque que toutes les trois tournent autour des mêmes thèmes, soient l'amour, les rêves, la nostalgie. Les paroles des chansons de Christophe et de Claude Barzotti parlent toutes deux d'amours déçues : la première parle d'une fille qui vie une peine d'amour et l'autre, d'un homme qui repense avec nostalgie à un amour d'été de son enfance. La chanson de Patrick Bruel raconte l'amour imaginaire qu'un homme se construit en croisant une femme dans un aéroport. Enfin, la chanson de Johnny Halliday parle d'un homme qui rencontre enfin la femme dont il rêvait depuis longtemps. On perçoit donc clairement une parenté entre ces chansons du point de vue thématique. La figure 5 présente une vue générale de la visualisation. Comme mentionné précédemment, la collection a été divisée en 5 *clusters*, lesquels sont représentés dans la visualisation par une couleur spécifique et ces mots-clés thématiques les plus fréquents. Le tableau 3 présente les mots-clés thématiques pour chacun des *clusters* :

Cluster 1 (mandarine)	aimer, amour, temps, jour, dire, vie, beau
Cluster 2 (jaune)	falloir, fille, monde, homme, nuit, soir, bleu
Cluster 3 (framboise)	croire, petit, mer, partir, revenir, souvenir, loin
Cluster 4 (vert foncé)	venir, nord, tourner, lune, danser, bon, pleurer
Cluster 5 (aqua)	vivre, mal, joie, falloir, peiner, attendre, perdre

Tableau 3. Mots-clés thématiques pour les 5 clusters

D'emblée, on remarque que le quatrième *cluster* semble réunir des chansons qui invitent à danser et à bouger avec des termes tels que « tourner » et « danser ». On y trouve entre autres les chansons « L'hôtesse de l'air » et « J'ai mis un tigre dans ma guitare », deux chansons à la fois drôles et entraînantes de Jacques Dutronc. Les deux premiers *clusters* semblent plutôt regrouper des chansons d'amour. Le premier *cluster* paraît réunir principalement des chansons positives (par exemple la chanson « La jeune fille du métro » de Renaud qui parle d'un coup de foudre qui se produit dans le métro parisien ou encore « La pêche à la ligne », toujours de Renaud, qui se termine mal mais qui commence par une déclaration d'amour). Au contraire, le deuxième *cluster* paraît rassembler des chansons nostalgiques (telles que la chanson « Mistral gagnant » qui aborde avec nostalgie l'amour d'un père pour sa fille) ou à propos d'amours qui se terminent (telles que « Comment te dire adieu » de Françoise Hardy). En revanche, les mots-clés thématiques « partir », « mer » et « loin » du troisième *cluster* suggèrent des chansons à propos de départs et de voyages. On y retrouve ainsi les chansons « Dès que le vent soufflera » et « Manhattan-Kaboul » de Renaud. Étant donné que notre corpus inclut des métadonnées riches sur les différentes chansons qui le composent, nous pouvons imaginer différentes possibilités pour le développement de fonctionnalités supplémentaires. Nous avons ainsi exploité la date de parution des chansons afin de permettre aux utilisateurs de comparer la distribution des chansons entre les *clusters* d'une période à

l'autre. Nous présentons dans le tableau 4 la répartition des chansons entre les *clusters* par période. Nous remarquons notamment des différences importantes entre les thématiques abordées dans les chansons des années 1950-1969 et celles des années 2000 à aujourd'hui. Par exemple, les thèmes du voyage ou des départs (*cluster* 3) semblent moins présents dans les chansons récentes. Dans les chansons des années 1950-1969, on remarque une forte proportion de chansons d'amour tristes (*cluster* 2) comparativement aux chansons d'amour positives (*cluster* 1).

	Clusters				
Périodes	1	2	3	4	5
Années 50-69	17,28	24.28	21,81	18,93	17,70
Années 70-79	22,47	19,59	20,41	19,79	17,73
Années 80-89	20,65	19,80	20,86	19,66	19,03
Années 90-99	20,35	20,39	19,42	20,04	19,80
Années 2000- 2013	20,26	20,38	18,93	20,16	20,26

Tableau 4. Répartition en pourcentage des clusters en fonction des périodes musicales

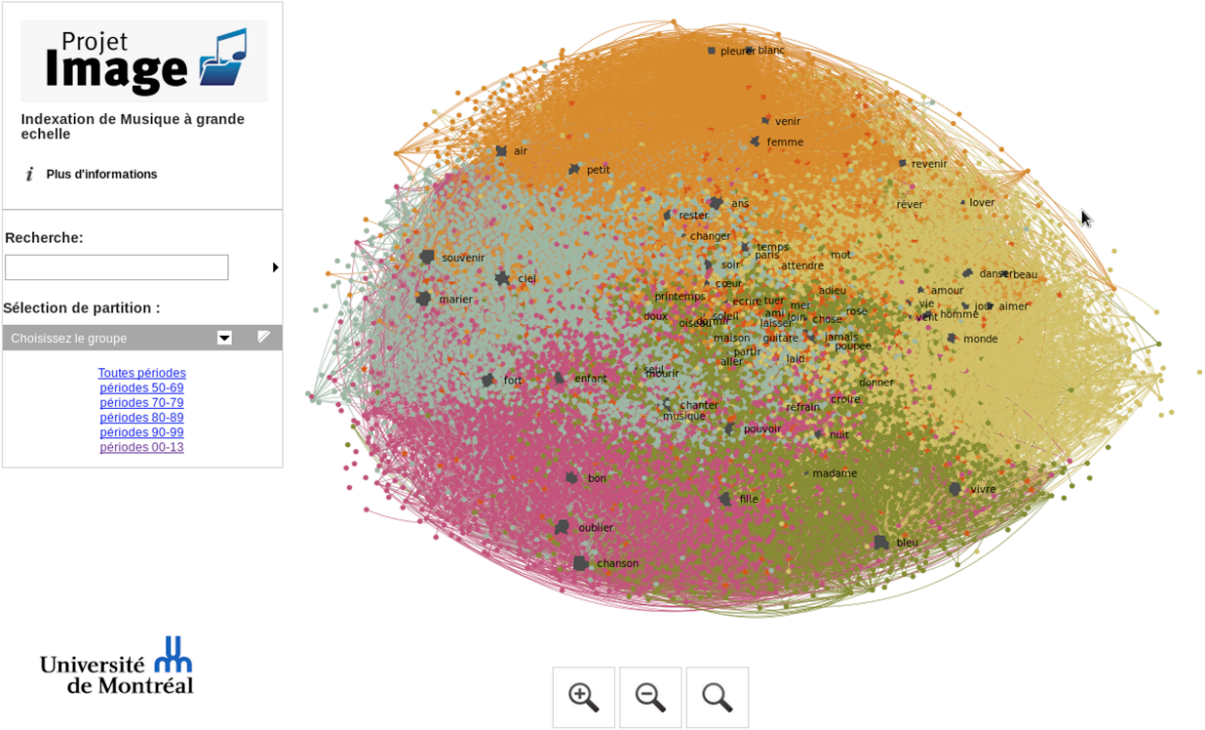

Figure 3. Vue d'ensemble de la collection.

7. Conclusion et perspectives

Cet article présente une méthodologie de fouille de textes intégrant une composante de visualisation de l'information afin d'extraire une structure thématique et de la représenter à partir d'une collection de chansons. Ce projet cherche à valider l'hypothèse selon laquelle l'analyse, l'organisation et la visualisation des paroles de chansons peuvent être utiles pour les utilisateurs de ce système d'information afin d'explorer, de parcourir et d'interpréter le contenu d'une grande collection musicale. Plus spécifiquement, le système repose d'abord sur

l'utilisation d'algorithmes de fouille de textes non supervisés et sur l'extraction automatique de mots clés thématiques en utilisant l'analyse sémantique latente. Le résultat obtenu prend la forme de regroupements de chansons caractérisés par des mots-clés thématiques qui en représentent le contenu. Ces informations sont transformées afin d'être représentées graphiquement dans une interface Web. Les applications de notre démarche sont nombreuses et variées dans le contexte du Web et de la diffusion de contenu multimédia (analyse de la proximité thématique d'artistes, analyse diachronique des thèmes et de l'influence musicale, etc.). Les chercheurs en musique populaire ou en sociologie sont les premiers utilisateurs ciblés par notre système. La disponibilité croissante de données musicales rend nécessaire le développement de ce type de système d'information. Notre système est l'objet d'un développement continu. L'utilisation de métadonnées supplémentaires permettra d'en enrichir les fonctionnalités. Ainsi, l'utilisation de métadonnées telle que l'année de parution, l'origine des artistes ou encore le genre musical pourrait permettre d'identifier d'éventuels liens entre des situations économiques et les thèmes abordés dans les chansons populaires et d'établir une corrélation entre les genres musicaux et les thèmes abordés. Finalement, en plus du développement des fonctionnalités de notre système, nous continuons à enrichir notre corpus et nous entendons procéder à une évaluation ergonomique de son interface.

Références

Auber D. (2003). Tulip : A huge graph visualisation framework. *In Graph Drawing Software, Springer Berlin Heidelberg, pp. 105-126.*

Bastian M., Heymann S. et Jacomy M. (2009). Gephi : An open source software for exploring and manipulating networks, in *International AAAI Conference on Weblogs and Social Media.*

Batagelj V. et Mrvar A. (1998). Pajek-program for large network analysis. *Connections, 21, pp. 47-57.*

Baur D., Steinmayr B. et Butz, A. (2010). SongWords: exploring music collections through lyrics. In Proc. of *ISMIR 2010,* pp. 531-536.

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

Bertin-Mahieux T., Ellis D.P., Whitman B. et Lamere P. (2011). The million song dataset. In Proc. of *ISMIR 2011*, October 24-28, 2011, Miami, Florida, pp. 591–596.

Béthune C. (2003). *Le Rap : une esthétique hors la loi*. Paris : Autrement.

Davies D.L. et Bouldin, D.W. (1979). A cluster separation measure. *Pattern Anal. Mach. Intell. IEEE Trans.* 2, pp. 224–227.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. et Harshman R. (1990). Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* 41, pp. 391–407.

Forest D. (2009). Impacts de la variation du nombre de traits discriminants sur la catégorisation des documents. *Actes Cinquième DÉfi Fouille Textes,* pp.77-83.

Geleijnse G. et Korst, J. (2006). Efficient lyrics extraction from the web. *In ISMIR.* pp. 371–372.

Grenier L. (1997). «Je me souviens»... en chansons : articulations de la citoyenneté culturelle et de l'identitaire dans le champ musical au Québec. *Sociologie et sociétés*, 29(2), pp. 31-47.

Hartigan J.A. et Wong M.A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), pp. 100–108.

Hu, X., Downie J.S. et Ehmann A.F. (2009). Lyric text mining in music mood classification. *Am. Music* 183, pp. 2–209.

Inskip C., MacFarlane A. et Rafferty P. (2010). Organising music for movies. *Aslib Proc.* 62, pp. 489–501.

Kessler R., Laplante A. et Forest D. (2014). Une interface visuelle pour l'exploration d'une collection de musique. In *14èmes Journées Extraction et Gestion des Connaissances* (EGC 2014), Rennes, pp. 347-352.

Kleedorfer F., Knees P. et Pohle T. (2008). Oh oh oh whoah! towards automatic topic detection in song lyrics. In Proc. of *ISMIR 2008,* pp. 287–292.

Knees P., Schedl M. et Widmer G. (2005). Multiple lyrics alignment: automatic retrieval of song lyrics. In Proc. of *ISMIR 2005,* pp. 564–569.

Laplante A. (2010). Users' relevance criteria in music retrieval in everyday life: an exploratory study. In Proc. of *ISMIR 2010,* pp. 601–606.

Laurier C., Grivolla J. et Herrera P. (2008). Multimodal music mood classification using audio and lyrics. In *Machine Learning and Applications 2008, ICMLA'08,* pp. 688–693.

Li T., Mitsunori O. et Tzanetakis G. (2012). *Music Data Mining. CRC Press Llc. ed.*

Logan, B., Kositsky A. et Moreno P. (2004). Semantic analysis of song lyrics. In ICME'04, IEEE International Conference, pp. 827–830.

Lübbers D. (2005). SoniXplorer: combining visualization and auralization for content-based exploration of music collections. In Proc. of *ISMIR 2005,* pp. 590–593.

Mayer R., Neumayer R. et Rauber A. (2008a). Combination of audio and lyrics features for genre classification in digital audio collections. In *Proceedings of the 16th ACM International Conference on Multimedia,* pp. 159–168.

Mayer R., Neumayer R. et Rauber A. (2008b). Rhyme and style features for musical genre classification by song lyrics. In Proc. of *ISMIR 2008,* pp. 337-342.

McKay C., Burgoyne J.A., Hockman J., Smith J.B., Vigliensoni G. et Fujinaga I. (2010). Evaluating the genre classification performance of lyrical features relative to audio, symbolic and cultural features. In In Proc. of *ISMIR 2010,* pp. 213-218.

Müller F. K., Damm D., Fremer C., Müller M. et Clausen M. (2007). Lyrics-Based Audio Retrieval and Multimodal Navigation. In *Music Collections - Springer* Vol. 4675, p. 112-123

Pampalk E. (2001). Islands of music: Analysis, organization, and visualization of music archives. Master's thesis Vienna Univ. Technol.

Pampalk E. et Goto M. (2006). MusicRainbow: A new user Interface to discover artists using audio-based similarity and web-based labeling. In Proc. of *ISMIR 2006 ,* pp. 367–370.

Rehurek R. et Sojka P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks,* Valletta, Malta, pp. 45–50.

Shuker R. (2013). *Understanding popular music culture.* 4e éd. Abingdon, Oxon : Routledge.

Van Zaanen M. et Kanters P. (2010). Automatic mood classification using tf* idf based on lyrics. In Proc. of *ISMIR 2010,* pp. 75-80.