Classification de documents Multimédia : vers une approche générale

Ismaïl Biskri, Louis Rompré, Lamri Laouamer & François Meunier

LAMIA – DMI - Université du Québec à Trois Rivières, Trois-Rivières (QC), Canada.

Abstract

In this article, we will present a generalization of numerical classification applied to the text, the image and the sound. The concept of n-grams, which since recent research, gives good results in the identification of the language, lexical clarification, complex terms learning or in the analysis of the analysis of speech, will be privileged to recognize the units of information describing the objects to be classified. This fact enables us to foresee an approach of classification less dependent on certain constraints connected to the support and the encoding of information.

Résumé

Dans cet article, nous présenterons une généralisation de la classification numérique appliquée au texte, à l'image et au son. La notion de n-grams, qui avec les recherches récentes, donne de bons résultats dans l'identification de la langue, la désambiguïsation lexicale, l'apprentissage des termes complexes ou dans l'analyse de l'oral, sera privilégiée pour reconnaître les unités d'information descriptrices des objets à classifier. Ce qui nous permet d'entrevoir une approche de classification moins dépendante de certaines contraintes reliées au support et à l'encodage de l'information.

Mots clés : classification, n-grams, texte, image, son.

1. Introduction

L'information de nos jours prend plusieurs formes. La langue, l'encodage, ou le type de l'information deviennent de plus en plus hétérogène. L'information textuelle, peu importe la langue de rédaction, est la plus répandue certes, cependant avec l'essor de l'Internet et des outils multimédias, celle-ci n'est plus la seule à véhiculer la connaissance. Le son et l'image prennent de plus en plus d'importance. Ne dit-on pas que dans certains cas une image vaut mille textes ? Cet état de fait a pour conséquence une nécessité de plus en plus perceptible de développer des outils à même de permettre de traiter l'image et le son, de les indexer, de les retrouver dans une base de données, de reconnaître leur forme, etc. (Goodrum, 2000) (Downie, 1999).

Aussi un rapide tour de l'état de l'art nous apprend, par exemple, que l'*indexation* fait actuellement l'objet de recherches très abondantes dans le domaine de l'analyse de texte, du traitement de l'image, de la vision par ordinateur ou de l'analyse du son. Il est proposé plusieurs méthodes pour associer à un texte, à une image ou à un son un ensemble de descripteurs de son contenu, dans le but de mesurer la ressemblance avec les descripteurs correspondant à une requête. *La reconnaissance* et la *classification* ne sont pas en reste. En effet plusieurs recherches leur ont été consacrés avec différentes techniques soit déterministes (symboliques) soit probabilistes (réseaux de neurones et classification) ou même encore des techniques basées sur des transformations mathématiques de morphologie tels que la

transformée de Fourier. Toutefois, malgré le nombre de ces applications potentielles, la complexité des algorithmes utilisés reste un facteur qui pose d'énormes problèmes. Il n'est pas évident de déterminer dans des temps raisonnables par exemple l'unité d'information qui va nous permettre de parcourir le contenu d'une image (Horng, 2002), ou le contenu d'un fichier son. Dans ces deux cas, plus on dispose de temps pour un traitement plus celui-ci a des chances de déboucher sur des résultats pertinents. Dans le cas contraire les résultats n'auraient aucun intérêt. Ainsi, par exemple, dans la détermination du contour d'une image par une approche polygonale, moins de temps consacré à cette tâche il y a, moins le contour s'identifiera à l'image. Et donc tout traitement fondé sur cette approche serait fatalement inintéressant. L'identification de l'unité d'information est également un problème pour la classification textuelle. En effet, même si traditionnellement, la « *tokenisation* » d'un texte se fait généralement par repérage des mots simples, ce processus ne permet pas une approche multilingue, étant donnée que certaines langues principalement germaniques ne disposent pas de délimiteurs de mots comme on en retrouve dans la langue Française (Manning & Schütze, 1999). Un problème analogue se pose pour le son. Celui-ci est un phénomène continu généré par une suite de surpressions et de dépressions de l'air par rapport à la pression atmosphérique. Ces vibrations forment un signal acoustique caractérisé par une amplitude[1], une fréquence[2] et une forme. Il est donc, par définition, un signal analogique. La majorité des sources sonores produisent des sons dont les vibrations sont de formes complexes composées de plusieurs sinusoïdes de fréquences et d'amplitudes différentes.

Dès lors une question simple s'impose : quelle est donc l'unité d'information atomique la plus adéquate pour segmenter un texte, une image, une onde sonore ? Balpe et al. (1996) soulignent, dans le cas du texte, que dépendant de l'objectif de lecture et de compréhension que nous nous donnons, la définition de l'unité d'information dépend de l'usage qui en est attendu. Dans une perspective de classification numérique de documents multimédias, la définition d'une unité d'information est tributaire des contraintes suivantes :

i. L'unité d'information doit être une portion du document multimédia soumis à l'analyse numérique ;
ii. Il doit être facile sur le plan informatique de repérer les unités d'information ;
iii. La définition d'une unité d'information doit être la plus indépendante possible de la complexité d'interprétation de l'encodage du document multimédia. Une telle définition permet à l'analyse numérique, moyennant des modifications minimes, de couvrir un large éventail de types de documents ;
iv. Les unités d'information doivent être statistiquement comparables. Il doit être aisé d'en calculer les fréquences d'apparition dans les différentes parties du texte et par conséquent d'estimer leur distribution et la régularité à laquelle plusieurs unités co-occurrent dans les mêmes parties du texte.

2. Les n-grams de caractères

Bien qu'ayant été proposée depuis longtemps et utilisée principalement en reconnaissance de la parole, la notion de **n-grams de caractères** prit davantage d'importance avec les travaux de Greffenstette (1995) sur l'identification de la langue, de Damashek (1995) sur le traitement de l'écrit. Entre autres, ils prouvèrent que ce découpage, bien que différent d'un découpage en mots, ne faisait pas perdre d'information. Parmi des applications plus récentes des n-grams on retrouve des travaux sur : l'indexation (Mayfield & McNamee, 1998) ; l'hypertextualisation

[1] L'amplitude détermine l'intensité du son.
[2] La fréquence détermine l'octave.

automatique multilingue avec les travaux de Halleb et Lelu (1998) qui, à travers une méthode de classification thématique de grandes collections de textes, indépendante du langage, construisent des interfaces de navigation hypertextuelle ; ou encore l'analyse exploratoire multidimensionnelle en vue d'une recherche d'information dans des corpus textuels (Lelu & *al*., 1998) et enfin dans le domaine de la désambiguïsation lexicale (Biskri & Meunier, 2002) et dans un système d'apprentissage des termes complexes (Biskri et al., 2004). Depuis récemment, un intérêt grandissant pour les n-grams est perçu dans la recherche en musicologie principalement en repérage de refrains (Patel & Mundur, 2005 ; Suyoto & Uitdenbogerd, 2005 ; Hsu & al., 2004), ou dans la recherche de similarité entre images (Laouamer & al, 2005).

Ainsi, dans le domaine textuel, on définira un n-gram de caractères par une suite de n caractères : bi-grams pour n=2, tri-grams pour n=3, quadri-grams pour n=4, etc. Par exemple le découpage en tri-grams du mot *informatique* donne inf, nfo, for, orm, rma, mat, ati, tiq, iqu, que.

Une image dans sa représentation bidimensionnelle (niveaux de gris) est une structure matricielle ou chaque élément représente l'intensité du pixel qui est comprise entre 0 et 255. Ainsi, un n-gram pour une image sera défini par une suite de valeurs d'intensité des pixels.

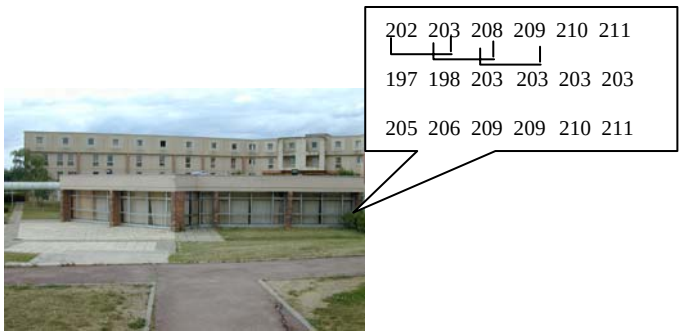

Figure 1: codage d'une image

Les n-grams de valeurs d'intensités de pixels pouvant être obtenus sont :
Les bi-grams : 202203, 203208, 208209, 20921, etc.
Les tri-grams : 202203208, 203208209, 208209210. etc.
Les quadri-grams : 202203208209, 203208209210, etc.

Un fichier sonore est une représentation numérique d'un signal acoustique. La conversion analogique/numérique consiste à effectuer un échantillonnage sur l'onde. L'objectif est de pouvoir reconstituer le signal à partir des échantillons prélevés. Lorsque l'on numérise un son, l'échantillonnage est effectué à un intervalle de temps régulier : la fréquence d'échantillonnage. Le son est alors représenté par un ensemble de valeurs d'échantillonnage qui peuvent être représentées sur 8 bits ou plus dépendant de la fidélité de numérisation souhaitée. Aussi, les unités d'information sont les n-grams de paire de valeurs amplitude-période associée à chaque mode du signal sonore redressé. Chaque paire amplitude-période représente dans les faits, la hauteur et la largeur du rectangle permettant l'approximation de l'aire sous la courbe du signal correspondant à une demi période du cycle de chaque événement sonore (note musicale) composant une trame monophonique complète (Figure 2).

Considérant ce qui précède, l'utilisation des n-grams apporte des avantages indéniables :

1. elle permet une production d'unités d'information pour toutes les langues utilisant un alphabet et la concaténation comme opérateur de construction de texte. Aucune

interprétation ou signification ne leur est associée. Il n'est plus nécessaire que reconnaître la langue soit un préalable utile. En outre, il est facile voire plausible de considérer une extension à des valeurs d'intensité de pixels qui forment des documents images ou à des valeurs d'échantillonnage du son qui forment des fichiers sonores ;

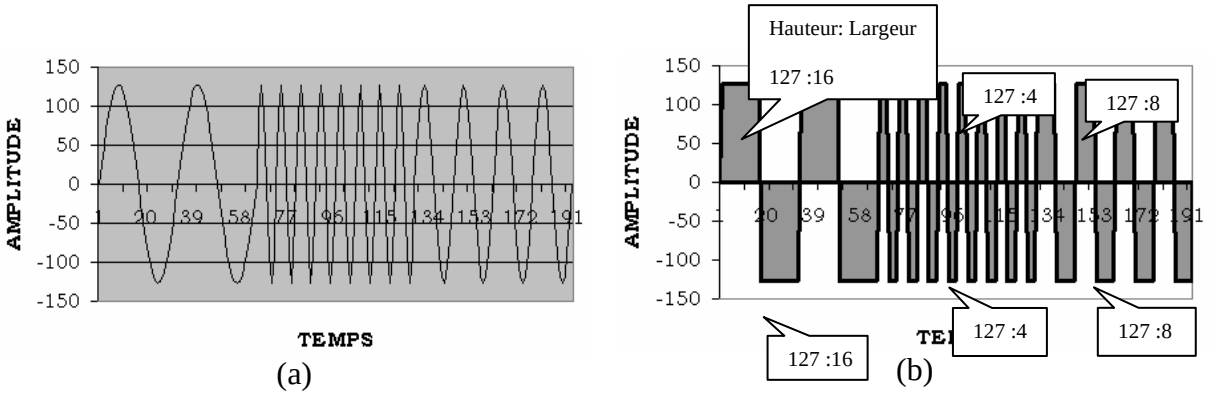

Figure 2 : Extraction des unités d'informations associées aux signaux sonores. (a) Onde sinusoïdale synthétisée. (b) Rectangles d'approximation correspondants et codage associé.

2. elle permet par exemple de contrôler dans le cas d'un texte la taille d'un lexique et le maintenir à un seuil raisonnable. La taille du lexique était jusqu'à présent l'aspect le plus controversé et considéré comme une limite des techniques fondées sur la comparaison des chaînes de caractères. En effet, un découpage en mots fait que la taille du lexique est d'autant plus grande que le corpus est grand. Cette limite subsiste malgré certains aménagements tels le "nettoyage" des mots fonctionnels, la lemmatisation et la suppression des hapax. Un lexique obtenu suite à un découpage en n-grams de caractères ne peut dépasser la taille de l'alphabet à la puissance n. Le choix d'un découpage en tri-grams pour une langue de 26 caractères donnerait une taille maximale de 26^3 entrées, soit un lexique de 17576 ti-grams possibles. Si on élimine les combinaisons qu'il est impossible de rencontrer (p.ex. AAA, BBB, CCC, HHA, etc.), ce nombre diminue de façon considérable ;
3. elle permet aussi de catégoriser de la même manière des expressions ou des textes qui ne sont que sensiblement identiques mais véhiculant la même sémantique comme dans les trois textes suivants : "l'informatisation de l'école", "informatiser l'école" et "introduire l'informatique à l'école". Ceci est également valable pour le son et l'image.

3. Les étapes de la classification

La classification numérique s'effectue au moyen d'un réseau de neurones ART comme celui utilisé dans (Meunier & al., 1997). L'unité d'information considérée est le n-gram de caractères (ici on utilise une appellation générique. Dans le cas de l'image le caractère représente la valeur d'intensité d'un pixel. Dans le cas du son le caractère représente la puissance du son pour le point d'échantillonnage), la valeur de n étant paramétrable. L'objectif visé est de prouver qu'il est possible de concevoir la même chaîne de traitement, peu importe la langue ou l'encodage de l'information, avec toutefois des aménagements dans la préparation des données à présenter au classificateur qui doivent être adaptés au type de la donnée (qu'elle soit textuelle, sonore ou de type image), ainsi qu'à l'objectif du traitement en particulier l'affichage des résultats. Notre approche n'est pas totalement automatique. Le

choix de certains paramètres est fait par l'utilisateur en fonction de ses propres objectifs et de sa subjectivité et de son rôle dans l'interprétation des résultats. Dans le cas d'une classification textuelle, l'input est un texte sous format ASCII. Dans le cas d'une classification appliquée aux images, l'input est un ensemble d'images au format JPEG. Dans le cas du son, l'input est formé d'un ensemble de fichiers sonores de type WAVE PCM mono. Le choix de ces configurations des données multimédias permet un traitement appliqué à des données brutes sans formatage particulier, mais aussi du fait qu'elles sont les plus répandues. Trois grandes étapes s'appliquent dès lors :

1. La **première étape** consiste à construire la liste des unités d'informations et des domaines d'informations. Les deux opérations se faisant simultanément, nous récupérons en sortie une matrice où seront répertoriés les fréquences d'apparition de chaque unité d'information dans chaque domaine d'information.
   a) Dans le cas du texte, les unités d'informations sont les n-grams de caractères. Les domaines d'informations sont représentés par des segments de texte dont on veut comparer la similarité. Le choix de la valeur du n du n-gram dépend de l'utilisateur et de l'expertise qu'il veut mener. Outre la valeur du n, d'autres paramètres sont considérés comme la possibilité de supprimer les caractères non alphanumériques en caractère espace, ou encore les chiffres. Ces deux paramètres répondent aux besoins d'une analyse pour laquelle les chiffres, la ponctuation ou encore d'autres caractères spécifiques seraient ou pas importants pour la qualité des résultats. Dans un texte technique par exemple, il serait peut être intéressant de savoir si version1 est différente de version2 et, par conséquent, les chiffres pourraient avoir autant d'impact informatif que les caractères alphabétiques. Un dernier paramètre concerne la conversion des caractères majuscules en minuscules pour distinguer ou non les lettres minuscules des majuscules. L'autre aspect important est le paramétrage de la segmentation qui peut se faire soit en des sections formées d'un nombre déterminé de phrases, de paragraphes ou de mots, ou tout simplement en des sections séparées par un caractère spécial. Ce paramètre est toujours choisi par l'utilisateur. Le lexique formé de n-grams de caractères subit au cours de cette première étape un nettoyage qui consiste en l'élimination des "n-grams hapax", ceux dont la fréquence est inférieure à un certain seuil ou supérieure à un autre seuil, ceux spécifiques sélectionnés dans la liste (par exemple des n-grams contenant des espaces) ou encore certains n-grams considérés comme fonctionnels, particulièrement les suffixes.
   b) Dans le cas de l'image, les unités d'informations sont les n-grams de valeurs d'intensités de pixels. Un domaine d'information est représenté par une image. La phase de pré-traitement consiste en plusieurs opérations qui s'appliquent directement à l'image. La première est représentée par la possibilité de conversion des images couleurs tridimensionnelles en des images aux « niveaux de gris » bidimensionnelles, pour permettre à l'ensemble des images d'avoir une dimension commune bidimensionnelle. Le lissage des images pour filtrer le bruit est une deuxième opération. Le bruit étant souvent de haute fréquence, il est donc possible en utilisant un filtre passe-bas de le réduire. Un inconvénient majeur de ce pré-traitement est de filtrer les hautes fréquences qui sont naturellement présentes dans l'image. Toutefois, nous pensons que ce filtre pourra éliminer les bruits induits par l'ajout de légendes textuelles sur l'image comme c'est très souvent le cas sur le WEB.
   c) En ce qui concerne le son, les unités d'information sont les n-grams de paire de valeurs amplitude-période tel que présenté dans la figure 2. Pour arriver à effectuer le codage des n-grams il faut au préalable effectuer certaines opérations de nettoyage

des données sonores. La première opération de nettoyage consiste à convertir l'amplitude des échantillons sur 8 bits lorsque ceux-ci sont codés sur 16 bits. Cette opération permet de limiter l'intervalle des valeurs d'amplitude à 256 niveaux ([-127, 127]) plutôt qu'à 65 536. L'étude de la conversion analogique/numérique a démontré qu'un même signal peut être quantifié de manière différente. Même si la qualité de la numérisation est affectée par la réduction de la quantification de chacun des échantillons, le signal peut tout de même être reproduit et reconnu. La deuxième opération de nettoyage consiste à lisser le signal sonore pour ainsi éliminer d'une part le bruit inhérent au processus de capture et d'autre part les harmoniques qui texturent les ondes de fréquences fondamentales. Dans le contexte de cette recherche, l'extraction des unités d'information sonores est exclusivement basée sur les fréquences fondamentales puisqu'elles offrent de par leurs plus grandes amplitudes un meilleur pouvoir discriminant des événements sonores. Le lissage est implémenté sous forme d'un filtrage passe-bas utilisant un noyau gaussien de la forme (Levine, 1985) :

$$g(i) = \frac{1}{2\pi\sigma_t^2}\exp\left[-\frac{i^2}{2\sigma_t^2}\right] \qquad (1)$$

où σ_t correspond à l'écart-type de la gaussienne dans le domaine temporel et permet de déduire la dimension du filtre numérique donnée par $4\sigma_t+1$. Le filtrage passe-bas gaussien est appliqué par une opération de convolution donnée par Gonzalez et al., 1987 :

$$s^*(t) = \sum_{i=-2\sigma}^{i=2\sigma} s(t)g(t+i) \qquad (2)$$

où $s(t)$ représente le signal sonore original et $s^*(t)$ sa version lissée.

La figure 3 présente les résultats de lissage pour différentes valeurs de σ_t, d'une onde sonore correspondant à une séquence de notes jouées au piano. À la lumière des résultats observés dans cette figure, il est intéressant de constater que plus la valeur de σ_t est petite (fig. 3 (a)) plus la courbe lissée (noir foncé) épouse l'onde sonore originale (gris). D'une façon opposée plus la valeur du σ_t augmente (fig. 3 (b) et (c)) plus l'onde sonore originale est lissée et comporte de moins en moins de fréquences élevées. Cette constatation peut facilement être expliquée et favorise la justification du choix d'une valeur de σ_t. D'abord, de par le théorème de convolution de la transformée de Fourier (TF) (Gonzalez et al., 1987), nous savons que la convolution dans le domaine temporel correspond à une multiplication dans le domaine fréquentiel. Dans le contexte du lissage gaussien cette correspondance s'exprime par :

$$s(t) * g(t) \Leftrightarrow S(u)G(u) \qquad (3)$$

où $S(u)$ et $G(u)$ correspondent aux TFs de $s(t)$ et de $g(t)$ respectivement. Le symbole * correspond à l'opérateur de convolution appliqué dans le domaine temporel. De plus, la forme de $G(u)$ est connue analytiquement :

$$G(u) = \exp\left[-\frac{u^2}{\sigma_u^2}\right] \qquad (4)$$

où σ_u représente la dimension du filtre gaussien dans le domaine spectral. En vertu du théorème de similarité de la TF (Gonzalez et al., 1987), le paramètre σ_u peut être

mis en correspondance avec son équivalent dans le domaine temporel σ_t par la relation :

$$\sigma_t = \frac{DIM}{2\pi\sigma_u} \qquad (5)$$

où *DIM* représente le nombre d'échantillons de signal sonore analysé.

Ce dernier résultat nous permet alors de justifier le choix de la valeur de σ_t. Sachant que nous voulons conserver par filtrage passe-bas les ondes fondamentales de fréquences inférieures à f_{max} et que σ_u peut être fixé à $f_{max}/1.2$, nous pouvons déduire à partir de l'équation (5) la forme de σ_t :

$$\sigma_t = \frac{1.2DIM}{2\pi f_{\max}} \qquad (6)$$

où le facteur $f_{max}= 1.2\ \sigma_u$ utilisé dans l'équation (6) correspond à la largeur à mi-hauteur à l'intérieur de laquelle l'amplitude de la gaussienne spectrale excède 0.5.

Par cette approche, le seul fait de connaître approximativement la fréquence fondamentale maximale de signaux sonores permettra de déduire le paramètre σ_t qui conditionne la dimension du filtre gaussien temporel utilisé pour effectuer un lissage préalable de ces signaux sonores.

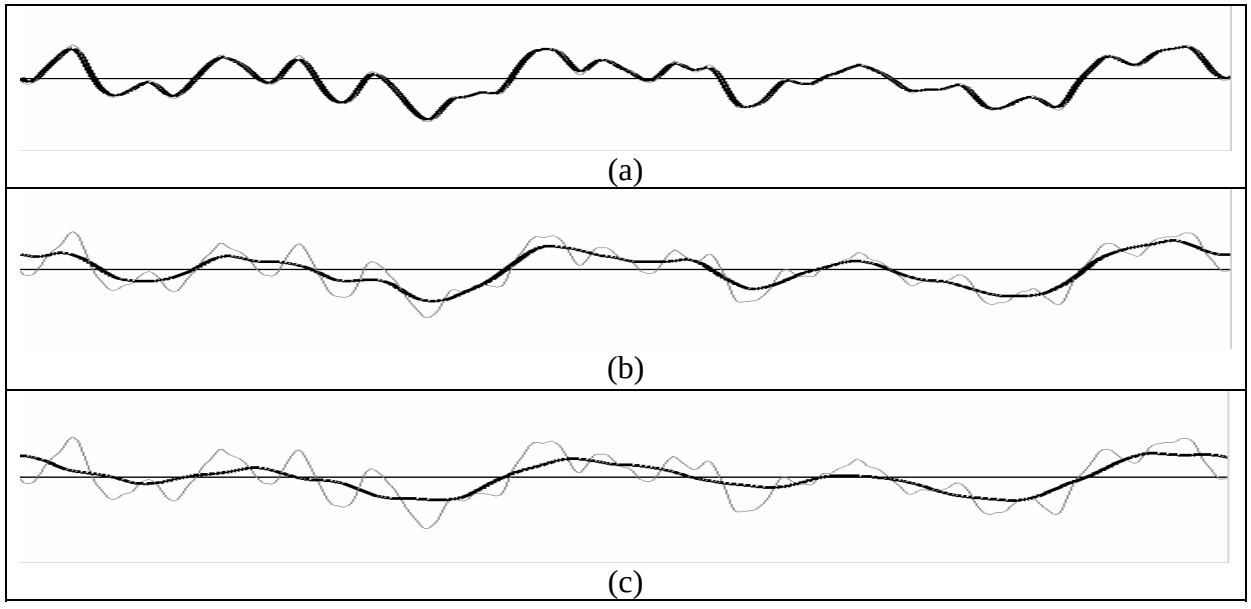

Figure 3 : Résultats de lissage. (a) σ_t=3.0. (b) σ_t=11.0. (c) σ_t=21.0.

2. Dans la **deuxième étape**, les segments représentés dans la matrice obtenue à l'étape précédente sont comparés entre eux au moyen d'un réseau de neurones (ART dans notre cas). Les domaines d'information semblables, étant donnée une certaine fonction de similarité, sont classés dans les mêmes groupes. En simplifiant, on peut dire que deux domaines d'information sont semblables s'ils sont constitués des mêmes n-grams avec des fréquences presque identiques. Le choix du réseau ART pour la classification n'est pas dicté par des raisons de performances particulières car tel n'est pas notre objectif pour le moment. Nous aurions tout aussi bien pu choisir un autre réseau neuronal qui aurait certes donné des résultats différents. De telles variations apparaissent dans les résultats des études expérimentales et comparatives des méthodes statistiques et numériques pour l'analyse de textes par ordinateur présentés dans la littérature.

3. La configuration du résultat de la classification numérique se présente par l'affichage des classes de segments et, pour chaque classe, l'affichage des domaines d'information qui la constituent d'une part, et du lexique (si possible) qui la forme d'autre part pour permettre l'interprétation des résultats par l'utilisateur. Cette étape peut aussi solliciter la saisie de propriétés particulières pour identifier les textes, les images ou les fichiers sonores en vue de les catégoriser pour d'éventuels traitements ultérieurs. Dépendant des paramètres choisis, les résultats d'une chaîne de traitement comme celle présentée ici peuvent servir à déterminer le thème principal de textes soumis à la classification (selon la perspective de l'utilisateur, son point de vue et sa connaissance du domaine), à déterminer l'acception d'un mot selon le contexte de son utilisation, à identifier des classes de similarités pour des images, ou encore pour des fichiers sonores. Ces résultats peuvent représenter un pré-traitement important dans certaines applications que ce soit la recherche d'information, le filtrage, la gestion des données multimédias, etc.

4. Évaluation

Nous avons mené plusieurs évaluations sur le texte et l'image. D'autres évaluations sont en cours actuellement particulièrement sur le son. Les évaluations sur le texte sont principalement de nature qualitative, alors que les évaluations sur les images et le son sont quantitatives. Nous n'avons pas comparé notre approche avec d'autres outils plus automatiques. Nous n'avons pas la même perspective étant donné que nous mettons la subjectivité de l'usager au centre du traitement.

a. Cas du texte :

La première évaluation a été réalisée sur un corpus formé d'une cinquantaine de pages (format ASCII) construit à partir d'extraits de documents trouvés sur le web. Ces documents couvrent divers domaines et permettent une hétérogénéité du contenu du corpus et, par conséquent, une meilleure compréhension des résultats de la classification. Pour les opérations préliminaires, soit la segmentation et l'extraction des n-grams, nous avons fixés les paramètres suivants : 10 phrases pour déterminer la taille d'un domaine d'information et 4 caractères pour déterminer la taille des n-grams. Nous avons considéré les lettres majuscules identiques aux lettres minuscules et nous avons remplacé les caractères non alphanumériques et les chiffres par des espaces. Nous avons ainsi récupéré 174 segments et 4 857 quadri-grams, après un "ménage" de la liste des n-grams qui a consisté à supprimer les n-grams contenant un ou plusieurs espaces et les n-grams ayant une fréquence égale à 1. La classification elle-même, au moyen du réseau de neurones ART avec un paramètre de vigilance de 0.1, donne lieu à la production de 100 classes de similarité, parmi lesquelles les classes suivantes :

- La classe 100 regroupe les segments 137 et 157. Le lexique de cette classe formé de l'intersection des lexiques des deux segments est constitué par {bourse, francs, marchés, millions, mobile, pdg, prix}. On constate, au regard de ce lexique, que le mot francs désigne la monnaie française et n'a aucun rapport avec la franchise ou avec les fameuses tribus "les francs". Ce même lexique nous renseigne également sur le thème commun que se partagent les segments 137 et 157, en l'occurrence le domaine financier.
- La classe 54 regroupe les segments 141 et 143. L'intersection des lexiques des segments de la classe 54 est formée de {appel, cour, décidé, juge}. Ainsi pour le mot cour, une seule signification est possible au regard des mots qui l'accompagnent cour de justice. On écarte aisément les sens suivants : la cour qu'on fait à une demoiselle, la cour de récréation, ou

encore les toilettes des Belges. Le thème de la classe 54 est par ailleurs bien identifié en ce sens qu'il s'agit de segments dont le contenu traite d'affaires judiciaires.

- La classe 98 regroupe les segments 71 et 73. Le lexique issu de l'intersection des lexiques des deux segments est formé des mots {culture, économiques, eurasistes, matérialiste, occident}. Dans ce contexte, le terme culture ne peut signifier que culture économique, et son utilisation n'est pas pour introduire une quelconque notion d'agriculture. Ce qui se confirme d'ailleurs avec le mot occident qui est utilisé ici dans le sens bloc géopolitique et non "là où se couche le soleil". Le thème de la classe 98 traite sans conteste d'options économiques ce que nous pouvons d'ailleurs vérifier au travers de la lecture des segments 71 et 73.
- La classe 64 regroupe les segments 166 et 167. Le lexique qu'on retiendra pour cette classe est formé de tous les mots dont la fréquence dans les segments 166 et 167 est supérieure ou égale à 2, en l'occurrence les mots {chance, dernière, dire, match, stade, supporters, vélodrome}. Le mot stade, du fait particulièrement de la présence des mots match, supporters et vélodrome, est compris comme étant un stade de football. Par ailleurs, pour un public averti qui sait que le vélodrome est le stade de Marseille, on comprend aisément que les deux segments 166 et 167 traite des supporters de l'Olympique Marseillais.
- La classe 13 regroupe les segments 32, 35, 41 et 48. Le lexique de cette classe formé de l'intersection des lexiques des quatre segments est constitué du seul mot russe. Celui-ci est suffisant pour nous permettre de conclure que le thème partagé par les quatre segments se rapporte à la Russie. L'union des lexiques, formée entre autres des mots conservateur, socialisme, marxiste, conservateur, révolutionnaire, Dostoievski, doctrine, impérial, slavophile, etc., nous permet de préciser que le thème de la classe 13 est dédié aux slavophiles et à la culture politique russe du 19^{ième} siècle. Une remarque s'impose : on imagine mal comment une classification fondée sur les mots aurait pu arriver à regrouper les segments 32, 35, 41 et 48 dans la même classe sans avoir recours à la lemmatisation étant donné que le seul mot commun est russe. Reste que la lemmatisation est relativement coûteuse en temps d'exécution et est une opération spécifique à chaque langue. Nous évitons ces inconvénients en utilisant les n-grams de caractères.

b. Cas de l'image :

La deuxième évaluation a porté, dans un premier temps, sur 23 images puis dans un deuxième temps, sur 52 images au format JPEG et de résolution 320*200. Certaines de ces images représentent le même objet mais sous différents angles. D'autres images représentent des objets de natures comparables (pas nécessairement le même objet) qu'il est possible de catégoriser de façon identique. Aussi, plusieurs résultats ont été observés.

Dans un premier test, nous avons considéré des bi-grams de valeurs d'intensités de pixels (BGVIP). Seulement les niveaux de gris ont été pris en compte. Deux variantes pour ce test ont été étudiées. Une première avec suppression des BGVIP hapax et une deuxième sans suppression des BGVIP hapax. Nous avons obtenu les résultats suivants :

Dans le premier cas, 11 classes ont été obtenues. 8 classes étaient légèrement bruitées alors que 3 classes contenaient une seule image et ne pouvaient donc être analysées.

Dans le second cas, 13 classes ont été obtenues. 5 classes étaient homogènes et pouvaient être considérées comme parfaitement cohérentes, 4 classes étaient légèrement bruitées et 2 classes étaient fortement bruitées. Les 2 dernières classes contenaient chacune une image et ne pouvaient donc être analysées.

Cet état de fait nous amène à conclure que la suppression des hapax dans le cas des images a un effet pervers sur le résultat de la classification. Contrairement à la classification textuelle où l'hapax est non porteur d'une capacité discriminante, celui-ci est très important dans le cas de la classification des images.

Dans un second test, nous avons maintenu un découpage en bi-grams BGVIP. Toutefois, nous avons considéré les images aux niveaux couleurs. 13 classes ont été obtenues, parmi lesquelles 9 sont parfaites. Les 4 autres contiennent chacune une image. Aucune classe bruitée n'a été obtenue. Il est possible, au regard de ce résultat de conclure que le fait de considérer des images couleurs tridimensionnelles est plus intéressant que si on les considérait aux niveaux de gris.

Nous résumons les résultats de ces deux premiers tests dans le tableau suivant :

	Classes		
	Parfaites	***Légèrement bruitées***	***Bruitées***
***Niveaux de gris Hapax supprimés* (11 classes)**	0	8	0
***Niveaux de gris Hapax non supprimés* (13 classes)**	5	4	2
Niveaux de couleurs (13 classes)	9	0	0

Un troisième test a été réalisé sur un ensemble de 52 images du format JPEG et d'une résolution 320*200. Deux cas ont été étudiés : le premier avec un découpage des images en bi-grams de valeurs d'intensité des pixels et un deuxième avec un découpage en tri-grams. Dans les deux cas, les images ont été considérées tridimensionnelles donc en couleur et les hapax n'ont pas été supprimés. Ce choix découle des résultats précédents.

Dans le premier cas 23 classes ont été obtenues parmi lesquelles 7 étaient absolument parfaites. Les images contenues dans ces classes étaient similaires. Donc pas de bruit dans ces classes. 8 autres classes étaient légèrement bruitées. Certaines contenaient une image qui ne devait pas être dans ces classes, d'autres ne contenaient pas une image qu'elles auraient du contenir. Toutefois, ces classes étaient suffisamment cohérentes pour être utilisables. Finalement, 3 classes étaient fortement bruitées.

Dans le second cas, nous avons obtenu 17 classes parmi lesquelles 6 étaient légèrement bruitées et 2 fortement bruitées. Les autres classes ne contenaient qu'une seule image chacune. C'est d'ailleurs cette dernière remarque qui nous amène à penser que le fait de découper les images en tri-grams augmente (peut être de façon exagérée) le taux discriminant. Nous résumons les résultats de ces deux tests dans le tableau suivant :

	Classes		
	Perfect	***Slightly noisy***	***Noisy***
***Découpage en bi-grams* (23 classes)**	7	8	3
***Découpage en tri-grams* (27 classes)**	0	6	2

c. Cas du son :

La troisième évaluation a porté sur la classification de 20 morceaux choisis dans deux partitions de Jean Sébastien Bach. Ils ont été mis dans 20 fichiers sonores différents. Les 10 morceaux de la première partition sont représentés par une séquence de notes situées entre les octaves 4 et 6 alors que les 10 morceaux de la deuxième partition sont représentés par une séquence de notes situées entre les octaves 3 et 5. Comme pré-traitement, l'onde sonore a été lissée à l'aide d'un filtre gaussien de paramètre sigma = 5.

Les résultats de la classification sont très intéressants. Celle-ci distingue les deux partitions. Les morceaux sont regroupés dans des classes différentes selon qu'ils appartiennent à la première partition ou à la deuxième. En effet, 7 classes représentent des regroupements de morceaux choisis dans la première partition alors que 3 classes regroupent des morceaux choisis dans la deuxième partition.

D'autres évaluations sont en cours et seront présentées dans nos prochaines publications.

5. Conclusion

Nous avons présenté ici une approche semi-automatique pour la classification de documents multimédias. Notre approche s'inscrit dans la continuité de nos travaux sur la classification textuelle. Nous justifions notre choix par le besoin d'avoir une approche générale pour classifier l'ensemble des sources de l'information, peu importe le type de leur encodage. Ce qui est très recevable étant données les contraintes auxquelles nous devons faire avec l'essor constant du web et les besoins qui en découlent. En effet, le concept du n-gram est indépendant de la nature de la donnée ou de l'interprétation sémantique qu'on peut lui associer. Il prend en compte, juste, le caractère (dans son sens large) utilisé pour représenter la donnée.

Nous avons montré l'adaptabilité de la notion du n-gram à l'image et au son et que la qualité des résultats est très significative. D'un point de vue computationnel, ces résultats confèrent à notre approche un coût (en terme de temps d'exécution) très bas.

Enfin, comme perspective de développements à venir, plusieurs voies sont possibles en particulier tester notre approche sur des documents multimédias complets qui contiennent et du texte et de l'image et du son.

Références

Balpe, J.P., Lelu, A. Papy, F. (1996). *Techniques avancées pour l'hypertexte*. Paris, Hermes.

Biskri, I. & Meunier, J.G. (2002). SATIM : Système d'Analyse et de Traitement de l'Information Multidimensionnelle. Actes du colloque JADT 2002, St-Malo, France, Mars 2002.

Damashek, M. (1995). Gauging Similarity with n-Grams: Language-Independent Categorization Of Text, *Science*, 267 : 843-848.

Downie J. S. (1999). Evaluating a Simple Approach to Musical Information Retrieval : Conceiving Melodic N-grams as Text, PhD thesis, University of Western, Ontario.

Goodrum A. A. (2000). Image Information Retrieval: An Overview of Current Research. *Information Science, Special Issue on Information Sciences*, Vol. 3, No. 2.

Gonzalez R. C., Wintz P. (1987). *Digital Image Processing*. Addison-Wesley.

Greffenstette. (1995). Comparing Two Language Identification Schemes. *Actes de JADT-95* : 85-96.

Halleb M., Lelu A. (1998). Hypertextualisation Automatique Multilingue à Partir des Fréquences de n-Grammes, *Actes de JADT-98*, Nice, France.

Horng, J. H. (2002). Improving fitting quality of polygonal approximation by using the dynamic programming technique. *In Pattern Recognition Letters*, v.23 n°14 : 1657-1673.

Hsu, J.L., Chen, A.L.P. & Chen, H.C. (2004). Finding Approximate Repeating Patterns from Sequence Data. In proceedings of the *5th International Conference on Music Information Retrieval*. Barcelona, Spain.

Lelu A., Halleb M., Delprat B. (1998). Recherche d'information et Cartographie dans des Corpus Textuels à Partir des Fréquences de n-Grammes, *Actes de JADT-98*, Nice, France.

Levine, M. D. (1985). *Vision in Man and Machine*. McGray-Hill Book Company.

Manning, C.D. Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Mayfield, J., & Mcnamee, P. (1998). Indexing Using both n-Grams and Words. *NIST Special Publication 500-242, TREC 7 :* 419-424.

Meunier, J.G., Biskri, I., Nault, G. & Nyongwa, M. (1997). Aladin et le Traitement Connexionniste de l'Analyse Terminologique. *Actes de RIAO-97*, Montréal, Canada : 661-664.

Patel N. & Mundur P. (2005). An n-gram based approach to finding the repeating patterns in musical. *in* Proc. Euro/IMSA 2005, Grindelwald, Switzerland.

Suyoto, I. S. H. & Uitdenbogerd, A. L. (2005). Simple efficient n-gram indexing for effective melody retrievial. In Proceedings of the *First Annual Music Information Retrieval Evaluation eXchange.*