Etude de textes par leur image

Hubert Marteau*, Alexandre Lefevre*, Nicole Vincent**

*Laboratoire d'Informatique, 64 av Jean Portalis, 37200 Tours
hubert.marteau@etu.univ-tours.fr
http://www.rfai.li.univ-tours.fr/RFAI/default.htm
**Laboratoire CRIP5-SIP, Université de Paris 5, 45 rue des Saints Pères,
75270 Paris Cedex 06
nicole.vincent@math-info.univ-paris5.fr
http://www.math-info.univ-paris5.fr/sip-lab/

Résumé. Nous proposons une méthode automatique de comparaison de textes reposant sur une technique de transformation d'un texte en une image de taille donnée et l'analyse à l'aide des outils de la géométrie fractale. Nous présentons une application à l'étude d'un corpus de 90 textes longs.

1. Introduction

De par la taille du corpus textuel existant et de par sa perpétuelle croissance, il est de plus en plus nécessaire de recourir à des outils de classification et d'indexation de textes. Nous proposons un système de représentation des textes lié à un système d'indexation tous les deux indépendants de la taille du corpus.

On peut distinguer trois types d'indexation : l'indexation manuelle, l'indexation semi-automatique (supervisée) et l'indexation automatique (non supervisée) [Pouliquen, 2002]. L'indexation manuelle nécessite qu'une personne désigne les termes d'indexation. Il est évident que ce type d'indexation n'est pas imaginable pour un passage à l'échelle. L'indexation semi-automatique propose à un utilisateur les termes d'indexation possibles selon leur fréquence, l'utilisateur n'a plus qu'à les accepter ou non. Enfin l'indexation automatique ne demande aucune participation de l'utilisateur. C'est évidemment ce dernier type d'indexation qui doit être mis en œuvre pour le traitement de gros corpus.

Nous proposons néanmoins, pour certaines applications, de faire reposer cette indexation automatique sur une phase d'apprentissage. Un des éléments, essentiel et dual de l'indexation, est la recherche qui repose sur la comparaison entre les représentations choisies pour l'indexation.

Il existe de nombreuses méthodes de mesure dans l'espace des unités textuelles et des mots [Lelu, 2002], nous pouvons en particulier citer le cosinus de Slaton [Salton, 1989], la distance du khi-deux, et le cosinus dans l'espace distributionnel.

Les méthodes d'indexation de textes se ramènent à un condensé d'informations des textes originaux : vecteurs, statistiques, par exemple. Le problème de cette nouvelle représentation est qu'elle n'admet aucune limite de taille prévisible. Les changements de représentations tels que les vecteurs, les statistiques, … permettent certes de condenser l'information, mais ne permettent pas d'obtenir une quantité fixe d'informations.

Afin de remédier à ce problème, nous proposons de transformer un texte en image, plus précisément d'associer une image de taille fixe à tout texte.

Cela assure une normalisation des textes et rend plus aisés les traitements et les comparaisons qui seront alors réalisés dans un espace de dimension finie.

D'autre part, après transformation du texte en image, nous pouvons imaginer que de nombreuses techniques développées dans le domaine du traitement d'images peuvent être appliquées. Enfin, cette méthode a l'avantage de pouvoir être appliquée aussi bien sur des textes assez longs comme des œuvres littéraires que sur des textes courts comme un entretien à questions ouvertes ou un discours politique. Cette méthode doit même pouvoir être appliquée à des textes « abstraits » comme un texte représentant un son [Dellandrea et *al*., 2002] ou un texte représentant une image [Nikolaou et Papamarkos, 2002].

2. Du texte à l'image

Contrairement au texte qui, dans sa structure même, comporte un aspect temporel très marqué, une image offre une sensation globale. Nous cherchons donc à transformer un texte en une image qui puisse mettre en évidence certaines structures.

Nous présentons dans cette partie une méthode de transformation d'un texte en image. Cette méthode n'est pas unique et cherche à s'appliquer sur tout type de texte.

1.1 Le texte

La construction de l'image se fait pixel par pixel et on peut dire que chaque pixel représente un motif textuel. Nous avons choisi de prendre les n-grammes de caractères comme unités textuelles. Un n-gramme est une suite de n caractères consécutifs. L'ensemble des n-grammes contenus dans un texte peut ainsi être listé en faisant glisser une fenêtre de n cases sur le texte étudié.

Jalam et Chauchat [Jalam et Chauchat, 2002] rappellent certains avantages que présentent les n-grammes : ils ne nécessitent pas de lemmatisation, on obtient directement la racine des mots sans prétraitement ; tolérance aux fautes d'orthographe et aux déformations éventuelles dans une limite raisonnable ; les n-grammes ne sont pas attachés à une langue particulière.

Les n-grammes ont surtout l'avantage de considérer un texte comme on considère une partition de musique, c'est à dire qu'une lettre est indépendante, quand on considère l'ensemble du texte, des lettres qui la précèdent. Le texte est donc considéré comme un bloc unique, inséparable.

Ainsi chaque pixel est représentatif d'un n-gramme, sachant que chaque n-gramme est, par sa fréquence d'apparition, représentatif d'une quantité d'informations contenue dans le texte original. Cette quantité d'informations est mise en valeur par le niveau de gris du motif qui lui est associé. Le motif représentant le n-gramme le plus fréquent est donc de couleur noire, et le motif représentant les n-grammes absents sont de couleur blanche. Pour l'ensemble des autres motifs, une échelle de niveaux de gris est construite en fonction de la fréquence du n-gramme le plus fréquent. L'ensemble des motifs est donc issu du texte d'origine.

Pour la construction des images, nous avons favorisé l'importance de la longueur des n-grammes. Pour créer une image carrée, ce qui conserve à l'image le plus de symétries possibles, nous avons, de plus, préféré prendre un alphabet ayant une cardinalité dont la racine carrée est entière. La langue française (et anglaise) utilise à la base 26 lettres, nous avons choisi de prendre un alphabet de 25 caractères.

Pour cela, nous avons regroupé les lettres « y » et « i » et les lettres « w » et « v ». Nous transformons les majuscules en minuscules et nous désaccentuons les lettres. Cela nous permet d'introduire un $25^{ème}$ caractère appelé caractère spécial, que nous représenterons par la suite par le caractère « & », ce dernier caractère remplace tout ce qui, dans un texte, n'est pas caractère littéral. Cette réduction a pour avantage de fixer une limite quant au nombre de n-grammes possibles : 25^n.

Chaque image est donc un ensemble de pixels dont le niveau de gris représente la fréquence d'apparition d'une suite de caractères recodés. Pour construire l'image il nous faut maintenant positionner les différents n-grammes dans un espace à deux dimensions. L'objectif est de mettre en évidence, grâce au principe de la localisation, les structures qui peuvent apparaître dans un texte.

1.2 L'image

Nous travaillons avec un corpus clos. Ce domaine d'application est très important surtout dans le cas d'un changement de représentation aussi radical que le nôtre. Pour que la méthode soit cohérente il faut qu'elle soit appliquée sur des images ayant au départ la même structure sous-jacente. Pour que cela soit possible, il faut que le patron (schéma d'organisation des motifs) soit le même pour toutes les images du corpus. Ce patron structure les images créées leur donnant ainsi un sens local et global.

Pour créer une image carrée avec un alphabet constitué de 25 caractères différents, il suffit d'organiser les motifs représentant ces caractères dans un tableau de pixels de taille 5x5. Afin que le patron soit représentatif du corpus, sa création dépend du contenu du corpus. Le motif représentant le caractère le plus fréquemment utilisé dans l'ensemble des textes est placé dans la case en haut à gauche du tableau et celui représentant le caractère le moins fréquent dans la case en bas à droite. Plus précisément, nous indiquons sur la figure 1 l'ordre utilisé pour le placement des motifs du plus fréquent au moins fréquent.

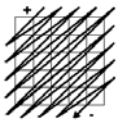

FIG. 1 - Organisation des motifs suivant leur fréquence d'apparition

&	E	S	N	L
A	I	U	C	V
T	O	P	Q	B
R	M	F	G	X
D	J	H	Z	K

FIG. 2 - Patron obtenu à partir des 1-grammes du corpus entier

FIG.3 - Image créée à partir des 1-grammes du texte : « Entretien n°1 »

Nous avons choisi, dans le cas d'un corpus clos, que le patron soit représentatif du corpus entier, donc une fois celui-ci adopté c'est le même patron qui sera appliqué sur l'ensemble des textes du corpus. On obtient ainsi une image par texte, et chaque image est construite en suivant une structure commune qui reflète une certaine logique dans la disposition des motifs par rapport à leur fréquence. Enfin, souhaitant une méthode indépendante de la langue, nous ne souhaitons pas fixer le patron a priori. Les figures 2 et 3 présentent le patron obtenu sur notre corpus d'application et l'image obtenue pour l'un des textes.

La création d'images à partir d'une lecture en n-grammes se fait naturellement en répétant à l'intérieur de chaque case, l'organisation établie précédemment. On s'approche ainsi de la construction d'une image fractale. Nous présentons en figure 4 la répétition du patron à une échelle d'observation plus précise.

Ainsi pour trouver l'information relative au motif « ab », il faut tout d'abord aller dans la case correspondant au caractère « a » (1ère colonne, 2ème ligne), puis dans cette case, il faut aller dans la case correspondant au caractère « b » (5ème colonne, 3ème ligne). Les coordonnées du motif « ab » dans le patron des 2-grammes de caractères (dimension : 25x25) est donc 5ème colonne, 8ème ligne. Nous présentons figure 5 l'image d'un texte codé avec les 2-grammes de caractères.

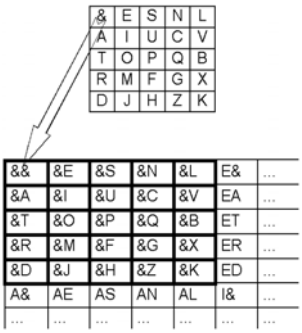

FIG. 4 - Récursivité du patron

FIG. 5 - Image créée à partir des 2-grammes du texte : « Entretien n°1 » (taille réelle 25x25 pixels)

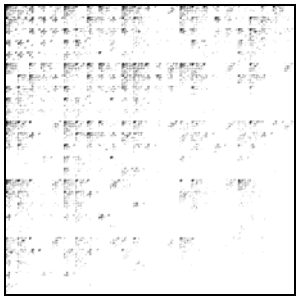

FIG. 6 - Image créée à partir des 4-grammes en utilisant une échelle deux fois logarithmique du texte : « Entretien n°1 » (taille réelle 625x625 pixels)

Avec une telle méthode de construction, les images construites à partir d'une lecture en 4-grammes ou plus paraissent entièrement blanches à cause de l'augmentation du nombre de motifs possibles et à l'échelonnage des niveaux de gris. Nous appliquons donc une double fonction logarithmique ce qui rééchelonne l'intensité des motifs, on obtient ainsi la figure 6. Cette dernière image représente de manière explicite une grande quantité d'informations. Elle fait apparaître la construction fractale.

La longueur optimale des n-grammes de caractères pour le type de discrimination que l'on souhaite réaliser, compte tenu du rapport entre l'amplitude de discrimination et le coût (taille de l'image : espace mémoire ; temps de calcul), a été fixé à 4. Cependant, pour une application nécessitant moins de précision comme la reconnaissance linguistique, cette longueur peut être raccourcie.

3. La méthode d'indexation

Cette méthode de construction est appliquée à l'ensemble des textes du corpus, il ne reste plus qu'à comparer les textes de manière à réaliser une classification des images entre elles, c'est-à-dire des textes entre eux. Par sa construction l'image révèle un caractère fractal. Nous avons donc choisi d'appliquer une méthode basée sur la dimension fractale de l'image comme méthode d'indexation.

La dimension fractale d'une image mesure la façon dont la fractale occupe l'espace de l'image, la dimension sera donc unique pour chaque image.

La notion de dimension d'autosimilarité se généralise au cas d'ensembles plus complexes par exemple en dimension de masse. A partir d'un point origine d'un ensemble X et en notant $m(X_k)$ la mesure de $X_k = X \cap [\ 0\ ;\ k\ ]^2$ l'expression de la dimension de masse est :

$$D = \lim_{k \to 0} \frac{\ln(\ m\ (X_k))}{\ln(\ k\ )}$$

Dans notre cas, des sous images apparaissent dans l'image, l'unité de base de chaque image est le point. La masse se calcule donc à partir des points et correspond à une mesure d'utilisation d'un type d'unité textuelle et de sa structuration. En effet, l'image est la représentation d'un texte et les points sont les représentations d'unités textuelles.

L'image paraît de type fractal, on peut calculer la masse sur l'image entière, on peut aussi considérer les sous images. Expérimentalement, on constate que l'évolution de la masse pour les sous-images est quasi-linéaire après transformation logarithmique. Le coefficient de régression de la droite formée par l'évolution des masses exprime la complexité de l'image, il correspond donc a une mesure d'autosimilarité représentative de l'image, nous avons pris cette valeur d'autosimilarité comme valeur d'indexation.

4. Application

Notre corpus d'application est un ensemble d'entretiens de type sociologique qui avaient pour but de déterminer le passage des valeurs au sein d'une même famille. Ces entretiens concernent 30 familles dans lesquelles ont été interrogés un enfant, la mère et le père, soit un total de 90 entretiens. Ces entretiens abordaient obligatoirement les thèmes suivants : la religion, l'éducation, l'autorité familiale, la politique et les travaux ménagers.

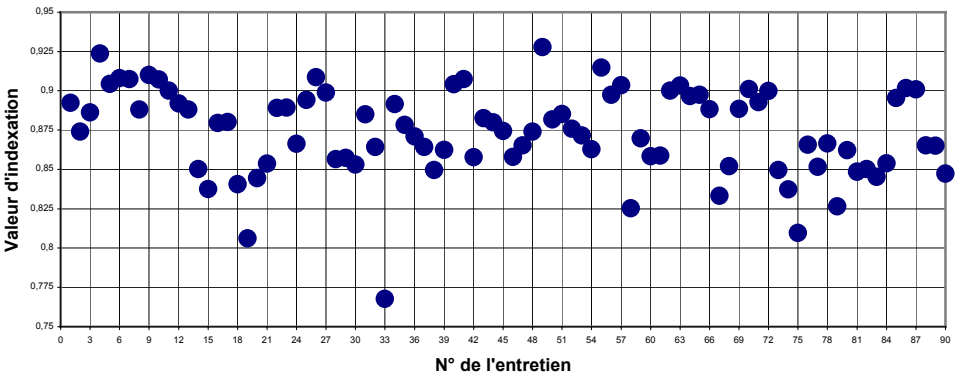

FIG. 7 - Valeur associée à chaque texte

La figure 7 présente le résultat obtenu par la méthode d'indexation pour chaque individu, les résultats sont organisés par famille et suivent, à l'intérieur d'une même famille, l'ordre énoncé précédemment : enfant, mère, père. L'entretien n°33, premier entretien remarquable, n'est qu'un résumé d'entretien, la structure de son image est donc bien moins complexe que la structure des images des autres entretiens.

Après analyse de l'ensemble des résultats on observe que les entretiens ayant une valeur assez élevée sont associés aux individus qui marquent une certaine indifférence quant aux sujets abordés, ils tiennent des discours plus complexes et ont tendance à s'éloigner de l'objectif principal des questions posées alors que les personnes attachées à certaines valeurs abordées dans les thèmes ont tendance à répondre plus directement aux questions.

5. Conclusion

Nous proposons ici une méthode qui permet de synthétiser l'information contenue dans des textes longs en transformant un texte en image. Cette transformation permet, de plus, l'application d'une méthode d'indexation graphique. Ce mode de représentation et d'indexation permet de distinguer les textes entre eux aussi bien par leur contenu que par leur forme. Nous avons porté l'application principale de notre étude sur des textes de type entretiens sociologiques, ce corpus peut néanmoins être agrandi aux textes littéraires en général, aux entretiens avec questions ouvertes ou aux textes abstraits ou codages, …

De même les méthodes de représentation et d'indexation étant indépendantes, elles peuvent être changées toutes les deux, on peut imaginer de nombreuses méthodes de représentation et on peut tester de très nombreuses méthodes de traitement d'images adaptées à la comparaison des représentations. Ainsi, cette méthodologie pourrait aboutir dans de nombreux problèmes liés aux textes comme l'identification d'auteurs ou la segmentation thématique par exemple.

Références

[Dellandrea et al., 2002] E. Dellandrea, P. Makris, M. Boiron et N. Vincent. A medical acoustic signal analysis method based on Zipf law, International Conference on Digital Signal Processing (DSP 2002), Vol. 2. p. 615-618, 2002.

[Jalam et Chauchat, 2002] R. Jalam et J. H. Chauchat. Pourquoi les n-grammes permettent de classer des textes ? Recherche de mots-clés pertinents à l'aide des n-grammes caractéristiques, JADT 2002.

[Lelu, 2002] A. Lelu. Evaluation de trois mesures de similarité utilisées en sciences de l'information, Les journées d'étude des systèmes d'information élaborée, 2002.

[Nikolaou et Papamarkos, 2002] N. Nikolaou et N. Papamarkos. Color image retrieval using a fractal signature extraction technique, Engineering Applications of Artificial Intelligence, Vol. 15, p. 81-96, 2002.

[Pouliquen, 2002] B. Pouliquen. Indexation de textes médicaux par extraction de concepts, et ses utilisations, Thèse de doctorat à l'Université de Rennes 1, 2002.

[Salton, 1989] G. Salton. Automatic Text Processing, Addison Wesley, 1989.

Summary

We propose a text comparison automatic method which transforms a text into an image of fixed size. We present an application to the study of a corpus of sociologic long texts which are compared together and for which conclusions on their contents can be deduced from the statistical study.