Classification non supervisée de documents hétérogènes: Application au corpus " 20 Newsgroups "

Julien Lemoine, Hamid Benhadda, Julien Ah-Pine

Classification non supervisée de documents hétérogènes: Application au corpus “20 Newsgroups”

Julien Lemoine
Thales
julien.lemoine
@fr.thalesgroup.com

Hamid Benhadda
Thales
hamid.benhadda
@fr.thalesgroup.com

Julien Ah-Pine
Thales
julien.ah-pine
@fr.thalesgroup.com

Abstract

Résumé : Cet article présente les résultats d'une méthode de classification non supervisée sur le corpus “20 newsgroups”. Les classes obtenues ont permis de retrouver les grandes thématiques du corpus ainsi que des sous-thématiques plus spécifiques.
Mots clés : 20 newsgroups, classification non supervisée, analyse relationnelle.

1 Introduction

Classifier de façon non supervisée un corpus donné, revient à y chercher des groupes ou classes de documents les plus homogènes[1] possibles. Nous proposons, dans cet article, une méthode de classification non supervisée qui s'appuie sur la théorie de l'analyse relationnelle. Cette théorie a été formalisée et développée par F. Marcotorchino et P. Michaud [1]. Parmi les principales qualités de cette théorie, on peut citer :

- la non fixation a priori du nombre de classes à trouver dans le corpus;
- sa robustesse vis à vis de la disparité en terme de longueur des documents à classifier;
- la stabilité des résultats obtenus vis à vis de l'ordre dans lequel les documents sont traités.

On appliquera, à titre illustratif, notre technique au corpus “20 Newsgroups” qui est constitué de 19,997 documents, issus de 20 forums différents et décrits par 145,980 descripteurs. Ce corpus est devenu une référence sur laquelle des techniques de Text Mining telles que la catégorisation ou la classification non supervisée sont testées et comparées. Sa caractéristique essentielle est son hétérogénéité en termes de taille des documents, en termes de thématiques et en termes de style.

2 Bases théoriques

2.1 Pré-traitements

Afin d'obtenir une représentation vectorielle en présence/absence des documents telle que celle proposée par G. Salton [2], on applique au corpus un traitement linguistique (étiquetage morpho-syntaxique) qui transforme un document en une liste de lemmes accompagnés de leur catégorie grammaticale (verbe, adverbe, nom commun, nom propre, préposition, etc. ...). A l'issue de cette étape d'étiquetage, la sélection des descripteurs qui participeront à la classification se fait en trois étapes :

1. On choisit les catégories grammaticales dont on veut tenir compte dans la classification[2]. Seuls les descripteurs dont

[1] Les thématiques traitées par deux documents d'une même classe doivent être plus proches que celles traitées par deux documents de classes différentes.

[2] Ce choix sera étroitement lié au corpus en présence

la catégorie grammaticale a été choisie seront retenus;

2. On élimine ensuite des descripteurs retenus à l'étape précédente tous les happax[3];

3. On applique enfin, aux descripteurs restants, un anti-dictionnaire afin d'éliminer ceux dont le pouvoir sémantique classificatoire est faible.

En utilisant les descripteurs retenus à l'issue de ces trois étapes, on construit une matrice rectangulaire R de présence/absence de taille $n \times p$ où n est le nombre de documents et p le nombre de descripteurs. La ligne i de la matrice R correspondra au document D_i représenté par le vecteur $\vec{D}_i$:
$\vec{D}_i = (d_{i1}, .., d_{ij}, .., d_{ip})$ où :

$$d_{ij} = \begin{cases} 1 & \text{Si le lemme } j \text{ est présent} \\ & \text{dans le document } D_i \\ 0 & \text{sinon} \end{cases}$$

2.2 Sélection des descripteurs dans le cadre d'un corpus hétérogène

L'étape de pré-traitement mentionnée ci-dessus s'avère insufisante pour obtenir un ensemble de descripteurs pertinents lorsqu'on désire classifier un corpus hétérogène. Afin de ne conserver que les descripteurs les plus informatifs, on choisit l'entropie comme mesure de descrimination. La formulation mathématique de l'entropie E_j d'un descripteur j est donnée par la relation suivante :

$$E_j = -\frac{d_{.j}}{n}\log_2(\frac{d_{.j}}{n}) - \frac{n-d_{.j}}{n}\log_2(\frac{n-d_{.j}}{n})$$

où :

$$d_{.j} = \sum_{i=1}^{n} d_{ij}$$

Cette mesure prend ses valeurs entre 0 et 1. Plus l'entropie d'un descripteur est proche de 1, plus le pouvoir de discrimination de ce dernier est grand. Ainsi, on décide de ne retenir, parmi les descripteurs obtenus après l'étape de pré-traitements, que les k premiers classés par entropie décroissante. Le choix de k dépendra de la répartition des valeurs des entropies.

[3]Descripteurs présents dans un seul document du corpus.

2.3 Critère d'agrégation

On définira l'accord entre deux documents comme le nombre de descripteurs qu'ils ont en commun. L'accord $A_{ii'}$ entre deux documents D_i et $D_{i'}$ se traduit mathématiquement par le produit scalaire de leurs vecteurs représentants $\vec{D}_i$ et $\vec{D}_{i'}$:

$$A_{ii'} = \vec{D}_i \bullet \vec{D}_{i'} = \sum_{j=1}^{p} d_{ij}d_{i'j}$$

L'accord A_{ii} du document D_i avec lui même est la "taille"[4] de ce document. On définit ensuite l'accord maximum possible $AM_{ii'}$ entre les documents D_i et $D_{i'}$ comme la plus petite des tailles[5] de ces deux documents.

$$AM_{ii'} = Min(A_{ii}, A_{i'i'})$$

Une fois ces quantités définies, on cherche à obtenir une relation d'équivalence representée par la matrice binaire X de taille $n \times n$ et de terme général $X_{ii'}$ tel que :

$$X_{ii'} = \begin{cases} 1 & \text{Si } D_i \text{ et } D_{i'} \text{ sont dans la même} \\ & \text{classe} \\ 0 & \text{sinon} \end{cases}$$

La matrice X recherchée maximise le critère mathématique d'agrégation $C(X)$ suivant :

$$C(X) = \sum_{i=1}^{n}\sum_{i'=1}^{n}(A_{ii'} - \alpha AM_{ii'})X_{ii'} \qquad (1)$$

où α est un paramètre qui représente le pourcentage de l'accord maximal possible jugé suffisant pour décider de mettre les documents D_i et $D_{i'}$ dans une même classe de la partition finale.

[4]Ce qui représente le nombre de descripteurs qui le décrivent et non sa longueur textuelle en terme de mots.

[5]En effet, il est évident que le plus grand nombre de descripteurs que deux documents peuvent partager ne peut pas dépasser la taille du plus petit de ces documents.

La règle de décison consiste à dire que deux documents D_i et $D_{i'}$ seront, a priori, dans la même classe dès lors que :

$$A_{ii'} - \alpha AM_{ii'} > 0 \quad (2)$$

La solution exacte de ce problème s'obtient par programmation linéaire en linéarisant les contraintes de représentation de X en tant que relation d'équivalence [1][3].

2.4 Choix du paramètre α

α est un paramètre qui appartient a l'intervalle $[0, 1]$ et qui peut être fixé arbitrairement par l'utilisateur ou déterminé automatiquement.

2.4.1 Choix arbitraire

Le choix de ce paramètre ne sera pas le même selon que l'utilisateur cherche plutôt des classes les plus homogènes possibles afin de découvrir des signaux marginaux ou qu'il cherche à déterminer les grandes tendances du corpus. Si l'utilisateur préfère gérer un nombre de classes réduit relativement facile à exploiter et à interpréter, il devra opter pour un paramètre α relativement petit. Si, en revanche, il préfère obtenir un résultat optimal d'un point de vue théorique et mathématique, qui donne des classes très homogènes mais en nombre important, il devra opter pour une valeur voisine de 0.5[6].

2.4.2 Calcul automatique

Une façon non arbitraire de définir le paramètre α est d'utiliser une fonction des accords et des accords maximaux de tous les couples de documents. Nous proposons la fonction suivante :

$$\bar{\alpha} = \frac{N}{D}$$

où

$$N = \sum_{i=1}^{n-1} \sum_{i'>i} A_{ii'}$$

$$D = \sum_{i=1}^{n-1} \sum_{\{i'>i/A_{ii'}>0\}} AM_{ii'}$$

N représente la somme de tous les accords entre deux documents différents.

D représente les accords maximaux des couples de documents différents ayant un accord strictement positif.

$\bar{\alpha}$ représente un degré de ressemblance entre les couples de documents ayant des accords strictement positifs.

2.5 Calculs des liens entre classes

Si, en théorie, les classes de la partition obtenue, à l'issue de la maximisation du critère (1), doivent être très distinctes les unes des autres, elles sont en pratique liées entre elles par un ensemble de descripteurs communs. On mesure l'intensité de cette liaison entre deux classes C_k et $C_{k'}$ par l'indice de similarité suivant :

$$S_{kk'} = \frac{\sum_{i \in C_k} \sum_{i' \in C_{k'}} A_{ii'}}{\sum_{i \in C_k} \sum_{i' \in C_{k'}} AM_{ii'}}$$

Grâce à cet indice, on présente pour chaque classe la liste des classes avec lesquelles elle possède un fort indice de similarité ainsi que les descripteurs qui ont permis d'établir ce lien par ordre d'importance décroissant. Ceci permettra à l'utilisateur d'avoir une vision plus large des grandes thématiques du corpus.

2.6 Descripteurs représentatifs d'une classe

Les classes seront représentées par les dix lemmes les plus discriminants. L'indicateur de discrimination I_j^C pour un lemme j dans une classe C, que l'on a adopté est donné par la formule :

$$I_j^C = \frac{\sum_{i \in C} d_{ij}}{\sqrt{d_{.j} \times |C|}}$$

où $|C|$ est le nombre de documents de la classe C.

[6] Au delà de 0.5, le résultat de la classification tendra vers la solution triviale qui consiste à mettre tous les documents isolés

3 Application à des corpus de grande taille

Dès lors que l'on souhaite traiter des corpus de grande taille, la résolution exacte du probleme (1) par programmation linéaire n'est pas envisageable en des temps raisonnables.

Pour cela, on utilise une heuristique [3] nous permettant d'obtenir une solution approchée du problème (1). On analysera les résultats de cette heuristique sur le corpus "20 Newsgroups".

3.1 Pré-traitements spécifiques au corpus "20 Newsgroups"

Avant l'étape de pré-traitements, le corpus "20 Newsgroups" est constitué de 19,997 documents décrits par 145,980 descripteurs. Après pré-traitements, le corpus résultant est constitué de 19,866[7] documents décrits par 17,460 descripteurs.Les pré-traitements effectués sur ce corpus sont les suivants :

- Suppression des entêtes "Usenet" dans les messages (expéditeur, destinataire, sujet, date, etc. ...);
- On ne retient que la catégorie "nom commun" après l'étiquetage morpho-syntaxique;
- Utilisation d'un anti-dictionnaire spécifique au corpus (contenant des noms sans valeur informationnelle pour le corpus, par exemple "people", "way", "thanks", ...).

3.2 Statistiques relatives à la similarité des documents

On définit la similarité entre deux documents D_i et $D_{i'}$ par :

$$\rho_{ii'} = \frac{A_{ii'}}{AM_{ii'}}$$

Comme le critère (1) cherche à mettre, a priori, dans une même classe les documents D_i et $D_{i'}$ qui vérifient (2), ceci est équivalent à $\rho_{ii'} > \alpha$.

Le nombre de couples de documents n'ayant aucun descripteur en commun[8] est 149,487,214 soit 76 % de tous les couples possibles. Ces documents ne partagent aucun vocabulaire, ils ne sont pas intéressants pour la suite de l'analyse, on se focalisera donc sur les 24 % de couples ayant une similarité positive.

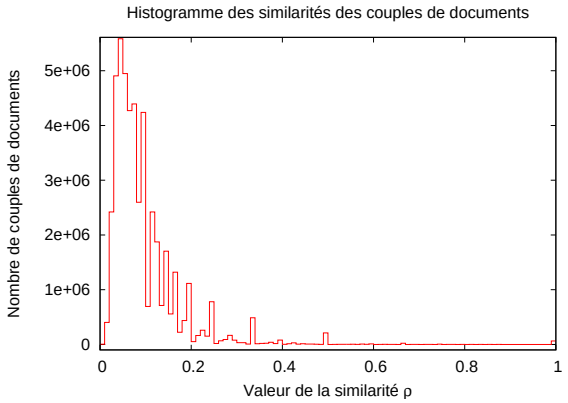

Figure 1: Nombre de couples de documents en fonction de la similarité

La figure (1) représente un histogramme avec en abscisse la similarité $\rho_{ii'}$ (par segment de 0.01) et en ordonnée le nombre de couples de documents ayant cette similarité.

Cette figure montre que 96.6% des couples de documents partagent moins de 25% de leur accord maximal.

Ces similarités, assez faibles entre les documents traduisent le fait que le corpus est très hétérogène.

3.3 Analyse du corpus

3.3.1 Sélection des descripteurs

En nous appuyant sur l'histogramme des valeurs de l'entropie des descripteurs du corpus donné par la figure (2), on constate que ces valeurs sont très petites. En effet, les entropies des 17,460 descripteurs sont comprises

[7] Les documents supprimés sont ceux qui n'ont plus de descripteurs après l'étape de pré-traitements

[8] c'est à dire ceux pour lesquels $\rho_{ii'} = 0$

entre 0.002 et 0.52. Ceci nous amène à dire que le corpus est difficilement classifiable.

Afin que les résultats de la classification ne soient pas trop bruités, nous avons choisi de ne retenir que les 1000 descripteurs[9] ayant les plus fortes entropies sachant qu'ils varient entre 0.05 et 0.52. Nous décidons donc de supprimer 16,460 descripteurs dont l'entropie est stritement inférieure à 0.05.

Malgré cette sélection, les descripteurs retenus restent peu discriminants. Néanmoins, comme nous le verrons par la suite, notre approche de classification permettra de dégager les grandes tendances au vu de ces descripteurs.

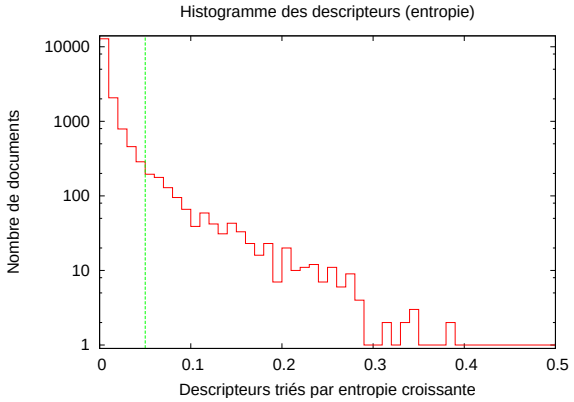

Figure 2: histogramme des valeurs de l'entropie des descripteurs

3.3.2 Calcul du paramètre $\bar{\alpha}$

Le paramètre automatique $\bar{\alpha}$ calculé est égal à 0.146. Après utilisation de notre heuristique, nous obtenons une partition composée de 330 classes dont la répartition de la taille des classes est donnée dans la figure (3).

Etant donnée la quantité d'informations à étudier, on analysera uniquement, ci-dessous, les classes dont le nombre de documents est supérieur à 100[10].

[9] une conséquence de cette sélection sera que 438 documents n'auront plus de descripteurs et seront donc retirés du corpus.

[10] La totalité du résultat est consultable à l'adresse suivante : http://clustering.speedblue.org

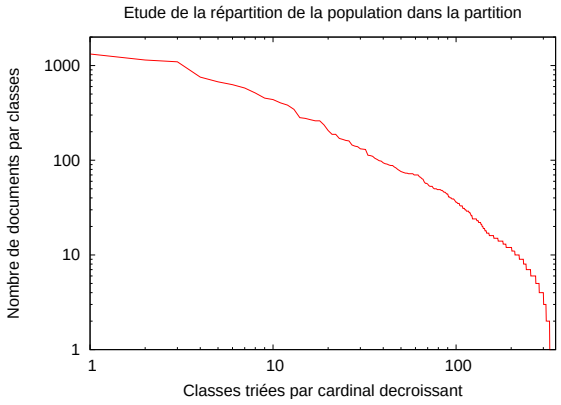

Figure 3: Répartition des documents dans les classes

3.3.3 Interprétation partielle du résultat

On donne dans la table (1) la liste des dix premières classes ainsi que leurs dix descripteurs les plus pertinents et leur cardinal (nombre de documents).

Cl	descripteurs	Card
1	game, team, player, hockey, season, playoff, fan, base ball, league, coach	1325
2	file, directory, program, window, ftp, FTP, archive, DOS, disk, server	1144
3	government, right, law, constitution, weapon, citizen, president, gun, policy, Turk	1095
4	car, engine, mile, tire, mileage, brake, dealer, wheel, auto, clutch	755
5	clipper, encryption, key, chip, escrow, crypto, wire tap, algorithm, privacy, government	673
6	drive, SCSI, IDE, disk, controller, ram, floppy, CD-ROM, jumper, soft ware	628
7	card, video, driver, ISA, monitor, bus, VGA, VLB, SVGA, graphics	579
8	religion, belief, god, Christian, atheism, atheist, faith, theist, truth, life	513

9	program, deficit, associate, cut, research, debate, scientist, tax, dollar, radio	453
10	window, application, manager, xterm, program, server, widget, display, screen, x11r5	437

Table 1: Dix première classes de la partition

Le calcul des indices de similarités entre les classes permet de déduire, lorsque ces indices sont forts[11], des thématiques plus globales. A titre d'exemple, l'indicide de similarité entre la classe 6 et la classe 7 est égal à 0.1, ce qui correspond à 71% du paramètre $\bar{\alpha}$. on donne ci-dessous (2) les descripteurs communs aux deux classes.

Classe 6	desc. liens	Classe 7
drive	card	card
SCSI	SCSI	video
IDE	IDE	driver
disk	controller	ISA
controller	disk	monitor
ram	monitor	bus
floppy	video	VGA
CD-ROM	ISA	VLB
jumper	ram	SVGA
soft ware	driver	graphics

Table 2: Lien entre la classe 6 et la classe 7

Nous donnons l'ensemble des thématiques globales dans la table (3) nous n'avons pas représenté les classes 11 et 34 car la sélection des descripteurs ayant été drastique, ces dernières sont peu significatives. La classe 11 contient des documents traitant de livres mais aussi bien de la vente de livre que de livres religieux. La classe 34 contient quant à elle des documents hétérogènes.

[11]On considère qu'un indice de similarité est fort lorsqu'il est proche du paramètre $\bar{\alpha}$

Thématique	Classes	Nb docs
Logiciel	2, 9, 10, 11, 24, 33	2716
Politique	3, 14, 15, 23, 27, 28, 31	2242
matériel informatique	6, 7, 17, 35	1923
Religion	8, 16, 25, 29, 32, 36	1318
Jeux	1, 26	1485
Cryptographie	5, 38	772
Automobile	4	755
Motocycles	12, 37	484
Vente	19, 20	442
Spatiale	21, 30	320
Medecine	18	261

Table 3: Principales thématiques de la partition obtenue

Certaines classes de la partition obtenue ont des liens forts et se regroupent autour de grandes thématiques. Toutefois, elles abordent des sujets particuliers de ces thématiques ce qui justifient le fait qu'elles soient isolées. Nous donnons les exemples suivants pour illustrer ces propos :

- la classe 6 de la thématique "matériel informatique", traite particulièrement des bus informatiques et du choix IDE ou SCSI;
- la classe 7 de la thématique "matériel informatique" aborde plus particulièrement les sujets relatifs au matériel video;
- la classe 14 de la thématique politique aborde particulièrement le problème des armes. Elle est décrite par "gun", "hand gun", "fire arm", "police", "weapon", "shot gun", "crime", "criminal", "accident", "cop";
- la classe 27 de la thématique politique traite en grande majorité des conflits israélo-palestiniens. Elle est décrite par "Jew", "Arab", "Israeli", "Palestinian", "holocaust", "land", "village", "peace", "inhabitant", "civilian";

- la classe 28 de la thématique politique qui traite des massacres/génocides (arméniens, bosniaques, kurdes, ...). Elle est décrite par “Turk”, “Armenian”, “extermination”, “mountain”, “genocide”, “road”, “soul”, “massacre”, “village”, “Muslim”.

Certaines classes obtenues contiennent des documents issus très majoritairement d'un forum, alors que d'autres contiennent des documents provenant de divers forums.

- la classe 4, traitant du sujet “automobile” provient majoritairement d'un forum :“rec.autos” 73.3%;

- la classe 7 sur le “matériel vidéo informatique” provient quant à elle de plusieurs forums : “comp.sys.ibm.pc.hardware” (31.4%), “comp.sys.mac.hardware” (17.7%), “comp.os.ms-windows.misc” (16.2%), “comp.graphics” (11.9%), “misc.forsale” (11.3%), “sci.electronics” (5.8%)

On est donc capable avec ces informations de tisser les interactions thématiques entre les différents forums.

4 Conclusion

Pour conclure, le travail que nous avons effectué ici nous a permis d'obtenir une classification du corpus “20 Newsgroup” sans échantillonnage.

Notre technique, au vu des résultats que nous avons obtenus, montre que nous avons pu mettre en évidence les grandes thématiques générales du corpus. Ces dernières ont été données dans la table (3).

Chacune de ces thématiques regroupe un ensemble de classes qui représentent des sous-thématiques plus spécifiques.

De plus, nous avons pu mettre en évidence des classes contenant des documents issus de plusieurs forums, identifiant ainsi des liens entre ceux-ci.

La théorie de la similarité régularisée[5] est une piste que nous comptons suivre pour améliorer de façon notable les résultats. En effet, elle permettrait, grâce aux pondérations fines tenant compte des structures internes des descripteurs qu'elle introduit, d'obtenir des résultats plus pertinents.

L'heuristique que nous avons utilisée ici fait partie d'un procédé qui a fait l'objet d'un dépôt de brevet.

References

[1] J.F. Marcotochino, P. Michaud (1981) Agrégation des similarités en classification automatique. Revue de Statistique appliquée, vol 30, No 2, 1981.

[2] G. Salton (1983). Introduction to Modern Information Retrival. McGraw-Hill, New York, 1983.

[3] J. Ah-Pine, H. Benhadda, J. Lemoine (2005). Un nouvel outil de classification non supervisé de documents pour la découverte de connaissances et la détection de signaux faibles : RaresText. Les systèmes d'information élaborée, Ile Rousse 2005.

[4] Y. Marom and I. Zukerman (2004). Improving Newsgroup Clustering by filtering author-specific words. Proceeding of PRICAI 2004. Auckland New Zealand.

[5] H. Benhadda (1998). La similarité régularisée et ses applications en classification automatique. Thése de doctorat de l'Université de Paris 6.

[6] J. Hartigan (1975). Clustering algorithms. John Wiley and Sons, New Yorks.

[7] T. Honkela, S. Kaski, K. Lagus, T. Kohonen (1996). Newsgroup Exploration with WEBSOM Method and Browsing Interface, Report A32, Helsinki University of Technology.