Clustering de documents dans des collections hétérogènes

Romaric Besançon* — **Anne-Laure Daquo***

* *CEA, LIST, Laboratoire Vision et Ingénierie des Contenus*
CEA Saclay Nano-INNOV PC n°173 91191 Gif-sur-Yvette CEDEX, France
{romaric.besancon,anne-laure.daquo}@cea.fr

RÉSUMÉ. *La classification non supervisée (ou* clustering*) de documents permet d'organiser thématiquement une collection de documents de façon à faciliter l'accès à l'information, ou à proposer une vue synthétique du contenu d'un ensemble de documents. Néanmoins, quand la collection considérée contient des documents de type différent, cette hétérogénéité perturbe les résultats du clustering, en regroupant plus volontiers les documents selon leur type que selon leur thème. Nous présentons dans cet article une approche simple pour la prise en compte de l'hétérogénéité de la collection dans le clustering, en utilisant une sélection des traits de représentation qui s'appuie sur les différences de distributions des termes selon les types de document. Nous montrons l'intérêt de l'approche proposée en proposant une évaluation sur un corpus hétérogène constitué spécifiquement pour cette tâche.*

ABSTRACT. *The goal of document clustering is to organize a collection of documents according to their topics, in order to facilitate the information access or to propose a synthetic view of the informational content of a collection of text. However, when the considered collection contains different types of documents, the clustering results tend to be impacted, because the similarity between the documents will rely as much on the type of the documents as on their topics. We present in this article a simple approach designed to take into account the type of documents in the document clustering task, using a feature selection method that exploits the type of the documents. We show the interest of this approach with an evaluation on a heterogeneous corpus specially designed for this task.*

MOTS-CLÉS : *Clustering de textes, hétérogénéité, sélection de traits.*

KEYWORDS: *Clustering, heterogeneity, feature selection.*

1. Introduction

Dans de grandes collections documentaires, le regroupement automatique des documents selon leurs similarités (clustering) facilite l'accès au contenu thématique, en faisant émerger des regroupements de documents similaires organisés idéalement autour d'un même thème. La plupart des recherches menées sur le clustering sont évaluées sur un unique jeu de données, une seule collection documentaire. Or, lorsque la collection de documents contient des documents de nature différente (par exemple des articles d'encyclopédie, des articles scientifiques, des brevets, des articles de presse, des tweets ou des pages web), des problèmes peuvent apparaître. En effet, les algorithmes de regroupement de documents s'appuient sur des mesures de similarité entre documents définies à partir des mots présents dans les documents. Plus les documents partagent un vocabulaire commun, plus ils sont considérés similaires et ont de chance d'appartenir à un même cluster. Or, des articles de presse et des notes de blog traitant d'un même sujet ne le feront pas forcément de la même manière, et les mots utilisés peuvent être plus représentatifs de la nature du document (genre documentaire, style rédactionnel, format, public visé) que de son thème. Si cette tendance est très marquée, on peut supposer que les techniques de regroupement de documents associeront de préférence dans un même cluster les documents de même nature et non de même contenu thématique, alors qu'une des caractéristiques généralement souhaitées, dans de nombreux cas d'utilisation du regroupement automatique, est de de proposer des regroupements documentaires transverses, indépendants du genre des documents.

Nous proposons dans cet article une approche pour limiter l'impact de l'hétérogénéité des collections documentaires sur le clustering, utilisant une méthode de sélection de traits pour la représentation des documents qui prend en compte le type des documents. L'idée est de sélectionner les traits les plus transverses aux différentes collections, en exploitant les métriques mesurant le pouvoir de discrimination des termes, habituellement utilisées en classification automatique. Dans ce contexte, ces mesures sont utiles pour retenir les termes les plus discriminants d'une classe ; dans notre cas, nous les utilisons au contraire pour supprimer les termes les plus spécifiques à un type particulier de documents. Dans la suite de cet article, nous présentons, en section 2, un état de l'art des techniques permettant de réduire l'influence des données hétérogènes sur le clustering et, en section 3, une présentation détaillée de l'approche que nous proposons. Enfin, nous rapportons, en section 4, les résultats d'expérimentations menées sur une collection hétérogène, en détaillant, d'une part, la construction de ce corpus et, d'autre part, les résultats de clustering obtenus.

2. État de l'art

La grande majorité des études sur le clustering évalue les différents algorithmes sur un unique jeu de données, contenant donc des documents de nature similaire, et ne traite pas le problème de l'hétérogénéité possible des données. Quelques études cependant se sont focalisées sur les jeux de données multi-collections.

(Marx *et al.*, 2003a ; Marx *et al.*, 2003b) proposent de traiter ce problème en introduisant la notion de type de document dans une variante du clustering fondée sur l'*Information Bottleneck* : ce type d'approche (Slonim et Tishby, 2000) s'appuie sur la construction de clusters de mots qui minimise l'information perdue sur les documents, puis sur la construction de clusters de documents qui minimise l'information perdue sur ces clusters de mots. (Marx *et al.*, 2003a) adapte cet algorithme pour créer un clustering couplé (*coupled clustering*) qui cherche à optimiser simultanément cette information et l'information de décentrage de la source. (Zhai *et al.*, 2004) s'intéressent à la fois aux thématiques partagées par les collections et à celles spécifiques à une collection : l'objectif est plus précisément de résumer pour chaque thème les similarités et dissimilarités des collections. Ils utilisent pour cela un modèle probabiliste génératif en introduisant des thématiques cross-collections et inter-collections dans le modèle. Dans la même lignée, des variantes des algorithmes de type *Topic Models*, et plus précisément LDA (*Latent Dirichlet Allocation*), ont également été proposées pour adapter ces approches à des données hétérogènes, en incluant par exemple des contraintes temporelles (Hong *et al.*, 2011) sur différents flux de documents ou en introduisant directement les sources des documents comme paramètres du modèle (Ghosh et Asur, 2013), dans un modèle nommé *Probabilistic Source LDA*, afin de créer des *topics* locaux à chaque source tout en maintenant une correspondance entre les *topics* des différentes sources. Parmi les méthodes plus simples, (Taralova *et al.*, 2011) proposent une variante du k-means fondée sur la définition de contraintes liées aux sources, qui obligent le k-means à regrouper des vecteurs candidats issus des multiples sources, en limitant le biais lié à une préférence pour une source. Les travaux ont été effectués pour le regroupement d'images, mais peuvent a priori être appliqués aux documents textuels.

De façon générale, ces différentes approches s'appuient sur des modifications d'un algorithme de clustering particulier, pour prendre en compte les différentes sources de documents directement dans la construction des clusters. L'approche que nous proposons est plus généraliste, en ce sens qu'elle ne dépend pas d'un algorithme de clustering en particulier : en modifiant l'espace de représentation des documents, on modifie l'étape de représentation vectorielle des documents. L'étape de clustering proprement dite est donc inchangée et on pourrait utiliser n'importe quel algorithme pour cette étape. De plus, dans ces différentes adaptations, certaines contraintes sont explicitement posées, comme par exemple le nombre minimal de sources présentes dans un même cluster pour (Taralova *et al.*, 2011), ou une indépendance explicite entre la distribution des termes dans les types de documents et leur distributions dans les clusters (Marx *et al.*, 2003a) alors que l'approche que nous proposons s'affranchit de ces contraintes. Enfin, certaines méthodes, comme celles fondées sur les *topics models*, même si elles permettent une intégration élégante de variables supplémentaires dans le modèle (comme la source des documents), sont en général relativement coûteuses et ne permettent pas toujours un passage à l'échelle, alors que la simplicité de la méthode que nous proposons la rend plus compétitive sur ce point.

3. Sélection de traits pour le clustering dans des collections hétérogènes

3.1. *Vue globale de l'approche*

Le clustering de documents s'appuie en général sur une représentation vectorielle des documents. De façon simple, cette représentation permet d'associer à chaque document un vecteur dont chaque composante est associée à une unité linguistique (un terme ou un concept), et la valeur de la composante représente l'importance de cette unité dans le document. Pour un terme du document, cette importance est en général fonction de la fréquence du terme dans le document et dans la collection, estimée selon une pondération de type *tf-idf* ou BM25. Certaines approches utilisent comme support de l'espace vectoriel des concept plus abstraits, implicites dans les approches de types *Latent Semantic Indexing* ou explicites (*Explicit Semantic Analysis*).

Sur la base de cette représentation vectorielle, l'algorithme de clustering exploite alors une similarité thématique entre les documents, calculée par une mesure de similarité entre les vecteurs, pour faire les rapprochements. Ces similarités sont calculées soit au cours de l'algorithme, soit au préalable, l'algorithme prenant donc en entrée, dans ce cas, une matrice de similarité contenant les similarités deux à deux entre les documents. Dans un contexte de grandes masses de données, cette seconde approche est néanmoins généralement coûteuse à mettre en œuvre.

Notre approche pour traiter le problème d'hétérogénéité des données se situe au niveau de la représentation des documents, en utilisant une sélection de traits orientée par la distribution des termes dans les différentes sous-collections correspondant aux différents types de documents. Cette approche permet donc d'être indépendant de l'algorithme de clustering utilisé pour le regroupement. Dans nos expériences présentées en section 4, nous avons utilisé un algorithme *k-means*. Pour une meilleure comparabilité, le k-means utilisé est initialisé avec la méthode KKZ (Katsavounidis *et al.*, 1994), qui peut être vue comme une version déterministe de kmeans++ dans laquelle chaque point initial est sélectionné comme le plus distant des points déjà choisis.

Par ailleurs, nous proposons une amélioration de cette approche par une méthode de *consensus clustering* (Vega-Pons et Ruiz-Shulcloper, 2011), qui permet de fusionner les résultats de plusieurs clustering, réalisés en s'appuyant sur différents paramétrages du clustering : dans notre cas, nous effectuons plusieurs clustering avec différents choix de sélection des traits. Cette fusion étant effectuée *a posteriori*, cette approche reste indépendante de l'algorithme de clustering choisi puisqu'elle n'utilise que des résultats de clustering. Nous présentons ces deux étapes plus en détail dans les sections suivantes.

3.2. *Prise en compte du type de documents dans la sélection des traits*

La sélection des unités de représentation influe sur la représentation vectorielle des documents et a fortiori sur les clusters. Pour un corpus hétérogène, le risque est alors de produire des clusters en lien direct avec l'une des sources documentaires, en parti-

culier si les collections se distinguent par des genres différents. Nous proposons donc de prendre en compte l'hétérogénéité des documents en sélectionnant les unités de représentation de façon à avoir des unités indépendantes de la nature des documents. Pour cette tâche, on peut s'appuyer sur les mesures utilisées en classification automatique de documents : dans ce cadre, ces mesures sont utilisées pour trouver les unités de représentation les plus caractéristiques des classes. Dans notre cas, à l'inverse, nous pouvons utiliser ces mesures pour filtrer les termes les plus spécifiques des collections documentaires, de sorte à ne garder que les unités les plus transverses.

Dans le domaine de la classification supervisée, de nombreuses mesures ont été proposées, dont on peut trouver des listes par exemple dans (Yang et Pedersen, 1997 ; Forman, 2003) ou (Manning *et al.*, 2008). Parmi ces mesures, les plus utilisées sont la fréquence en documents, la mesure du χ^2, l'information mutuelle ou le gain d'information (*Information Gain*). D'autres mesures, comme le *log-likelihood ratio* proviennent plutôt des analyses comparatives de distributions de vocabulaires issus de différents corpus (Dunning, 1993). La plupart de ces mesures peuvent être calculées à partir d'une table de contingence indiquant, pour chaque terme, le nombre de documents contenant ou ne contenant pas ce terme, selon que ces documents sont ou ne sont pas dans la classe considérée. Ces mesures sont présentées dans le tableau 1.

D'autres mesures, que nous ne considérons pas ici, ont également été proposées, comme par exemple la séparation bi-normale (Forman, 2003) ou des mesures de bootstrap (Lijffijt *et al.*, 2011). Dans tous les cas, l'idée de ces mesures est de sélectionner les termes les plus représentatifs d'un ensemble donné de documents (une classe, un corpus, un type de document) par rapport aux autres. Dans notre cas, regrouper des documents hétérogènes provenant de plusieurs collections selon des thématiques transversales aux différentes collections peut être facilité par un filtrage de ces termes les plus représentatifs. Privilégier ainsi les termes transverses permet aussi de réduire les variations inter-collections et de limiter, par exemple, l'influence d'un genre documentaire au sein même d'une collection.

Plus précisément, on note $spec(t_i, c_j)$ le score mesurant la spécificité du terme t_i dans la sous-collection c_j correspondant à un type de document particulier, calculée selon une des mesures du tableau 1. Dans une collection contenant plus de deux types de documents, cette spécificité est calculée entre cette sous-collection et les documents de tous les autres types cumulés. Pour calculer le score de spécificité global du terme, on prend le score maximum sur tous les types de documents :

$$spec(t_i) = \max_j spec(t_i, c_j)$$

Pour la sélection, on trie les termes selon leur score $spec(t_i)$ et on élimine les k termes avec les scores les plus élevés (*i.e.* les plus spécifiques), k étant un paramètre de cette approche, dont nous avons testé plusieurs valeurs dans notre évaluation.

Notations

Table de contingence

	doc $\in c_j$	doc $\notin c_j$	
doc $\ni t_i$	$f(t_i, c_j)$	$f(t_i, \bar{c_j})$	
doc $\not\ni t_i$	$f(\bar{t_i}, c_j)$	$f(\bar{t_i}, \bar{c_j})$	
			N

$$f(t_i) = f(t_i, c_j) + f(t_i, \bar{c_j})$$
$$f(\bar{t_i}) = f(\bar{t_i}, c_j) + f(\bar{t_i}, \bar{c_j})$$
$$f(c_j) = f(t_i, c_j) + f(\bar{t_i}, c_j)$$
$$f(\bar{c_j}) = f(t_i, \bar{c_j}) + f(\bar{t_i}, \bar{c_j})$$

Mesure	**Définition**
Fréquence en documents	$df(t_i, c_j) = f(t_i, c_j)$
χ^2	$chi2(t_i, c_j) = \sum_{\substack{t \in t_i, \bar{t_i} \\ c \in c_j, \bar{c_j}}} \frac{(f(t,c) - f(t)f(c)/N)^2}{f(t)f(c)/N}$
Information Mutuelle	$MI(t_i, c_j) = \sum_{\substack{t \in t_i, \bar{t_i} \\ c \in c_j, \bar{c_j}}} \frac{f(t,c)}{N} \log \frac{Nf(t,c)}{f(t), f(c)}$
Gain d'information	$IG(t_i, c_j) = e(f(c_j), f(\bar{c_j})) - \sum_{t \in t_i, \bar{t_i}} \frac{f(t)}{N} e(f(t, c_j), f(t, \bar{c_j}))$ avec $e(x, y) = -\frac{x}{x+y} \log \frac{x}{x+y} - \frac{y}{x+y} \log \frac{y}{x+y}$
Log-Likelihood Ratio	$LLR(t_i, c_j) = 2 \sum_{c \in c_j, \bar{c_j}} f(t_i, c) \log \frac{f(t_i, c)}{f(\bar{t_i}, c) \frac{f(t_i)}{f(\bar{t_i})}}$
	NB : pour LLR, les $f(.)$ sont des fréquences en occurrences et non des fréquences en documents

Tableau 1. *Différentes mesures pour la spécificité d'un terme t_i dans une sous-collection c_j correspondant à un type de documents.*

3.3. *Consensus clustering*

Conjointement à cette approche par réduction de l'espace de représentation aux termes les moins spécifiques aux types des documents, nous étudions également une approche pour limiter l'impact de l'hétérogénéité des collections fondée sur la fusion de plusieurs résultats de clustering. L'idée est d'effectuer plusieurs regroupements sur le même jeu de données, en faisant varier les unités de représentation, puis de

fusionner les résultats obtenus par ces différents regroupements. La combinaisons de différents résultats devrait alors permettre d'avoir une solution consolidée et, on peut l'espérer, indépendante de la source des documents.

Dans cette optique, nous nous intéressons aux approches générales de fusion de résultats de clustering, dans l'optique du *consensus clustering* (ou *clustering ensemble*) (Vega-Pons et Ruiz-Shulcloper, 2011). De façon générale, le *consensus clustering* est souvent utilisé pour améliorer la qualité globale du regroupement et/ou acquérir de la robustesse. Il réconcilie des propositions de regroupement de données qui peuvent varier selon le choix des algorithmes de clustering eux-mêmes ou leurs paramètres (par exemple : l'initialisation des centroids de manière aléatoire, le choix de la mesure de distance utilisée) ou le découpage du jeu de données. Selon ces variations, les objets attribués à chaque cluster ne sont pas toujours les mêmes, les formes des clusters peuvent d'ailleurs fortement varier et le nombre même de clusters peut lui aussi changer selon les partitions. Même si les différentes études sur le consensus clustering n'évoquent pas son utilisation directe pour regrouper des documents issus d'une collection hétérogène, un des cas d'utilisation évoqués par (Strehl et Ghosh, 2003) propose de combiner des propositions de regroupement de données sur la totalité de la collection mais utilisant des unités de représentation différentes. Couplé à une méthode de sélection des unités de représentation correspondant aux différentes sources, on propose alors de créer plusieurs regroupements et de les fusionner à l'aide des nombreuses méthodes de consensus clustering (Ghosh et Acharya, 2013), qui peuvent en particulier être probabilistes ou fondées sur des matrices de co-association entre documents. Dans cet étude, nous avons privilégié une méthode simple et robuste, en appariant les différents regroupements de données obtenus par clustering grâce à l'algorithme d'affectation de Khun-Munkres (Kuhn, 1955) (aussi appelé *algorithme hongrois*). Le premier résultat de clustering est considéré comme pivot et sert de base pour les appariements. On cherche pour chaque nouveau résultat de clustering à coupler les clusters à ceux du clustering pivot. Nous avons utilisé une méthode de consensus par vote (suivant la règle de la majorité) : un document est attribué à un cluster si les différents résultats de clustering l'ont associé majoritairement à ce cluster. Dans le cas contraire, le système prend un choix par défaut, c'est-à-dire la proposition faite par le clustering pivot.

4. Évaluation

4.1. *Mesures d'évaluation*

Pour l'évaluation des résultats de clustering par rapport à un classement de référence, de nombreuses mesures existent, parmi lesquelles les mesures classiques de précision/rappel s'appuyant sur les paires de documents, selon qu'ils appartiennent ou non au même cluster dans le résultat et dans la référence. On compte aussi la pureté, l'entropie, ou l'information mutuelle normalisée (NMI pour *Normalized Mutual Information*), définie par la formule suivante (Manning *et al.*, 2008), en notant

$\Omega = \{\omega_1, \ldots, \omega_J\}$ l'ensemble des clusters et $C = \{c_1, \ldots, c_K\}$ l'ensemble des classes de référence :

$$\text{NMI}(\Omega, C) = \frac{I(\Omega, C)}{[H(\Omega) + H(C)]/2} \quad \text{avec} \begin{cases} I(\Omega, C) = \sum_{\omega_i \in \Omega} \sum_{c_j \in C} \frac{|\omega_i \cap c_j|}{N} \log \frac{|\omega_i \cap c_j|}{|\omega_i||c_j|} \\ H(\Omega) = -\sum_{\omega_i \in \Omega} \frac{|\omega_i|}{N} \log \frac{|\omega_i|}{N} \end{cases}$$

Nous considérons également aussi les mesures de la famille BCubed (Amigó *et al.*, 2009), dans lesquelles les scores de précision/rappel sont calculés pour chaque document, plutôt que pour chaque paire de document. Plus précisément, pour un ensemble de documents $D = \{d_1, \ldots, d_N\}$, si on note $\Omega(d_i)$ (resp. $C(d_i)$) le cluster (resp. la classe de référence) contenant le document d_i et correct(d_i, d_j) une fonction qui vaut 1 si $\Omega(d_i) = \Omega(d_j) \Leftrightarrow C(d_i) = C(d_j)$ et 0 sinon, alors les mesures de précision et rappel BCubed sont définies par :

$$\begin{aligned} \text{precision-bcubed} &= \frac{1}{N} \sum_{d_i \in D} \frac{1}{|\Omega(d_i)| - 1} \sum_{d_j \in \Omega(d_i)} \text{correct}(d_i, d_j) \\ \text{rappel-bcubed} &= \frac{1}{N} \sum_{d_i \in D} \frac{1}{|C(d_i)| - 1} \sum_{d_j \in C(d_i)} \text{correct}(d_i, d_j) \end{aligned} \quad [1]$$

4.2. *Corpus*

A notre connaissance, il n'existe pas dans l'état de l'art de *benchmark* de référence pour cette tâche : chaque évaluation est faite de façon spécifique sur un corpus particulier. Nous proposons ici la construction d'un corpus en anglais, à grande échelle, créé spécifiquement pour cette évaluation. Ce corpus s'appuie sur la combinaison de documents de trois sources différentes : des articles de presse provenant du corpus Reuters-RCV1 (Lewis *et al.*, 2004), des pages encyclopédiques de Wikipédia et des pages Web issus du répertoire DMOZ *Open Directory Project*[1]. Les documents contiennent au moins 50 mots. Le texte des pages Wikipédia a été extrait en utilisant l'outil WikiExtractor[2]. Un filtrage par le contenu est appliqué aux documents Wikipédia en particulier pour éliminer les portails, les pages administratives, de téléchargement de logo, etc. Quant aux documents DMOZ, les pages dont la langue utilisée n'est probablement pas l'anglais sont éliminées. Les pages Web ont été nettoyées avec l'outil Tika [3] pour supprimer les éléments liés au format des documents (HTML).

Les collections Wikipedia, Reuters et DMOZ sont alignées. Cet alignement repose sur les annotation des documents en thèmes, qui sont partagées par les trois collections. Plus précisément, nous sommes partis des indications de catégories du corpus

1. http ://www.dmoz.org
2. http ://medialab.di.unipi.it/wiki/Wikipedia_Extractor
3. http ://tika.apache.org/

Reuters, qui a souvent été utilisé pour évaluer des résultats de clustering : chaque article est annoté avec des codes de catégories, organisés en catégories thématiques (*topic*), géographique (*region*) ou industrielle (*industry*). La catégorie *topic* concerne la thématique générale de la dépêche (sports, société, économie) alors que la catégorie *industry* traduit les domaines industriels concernés par la dépêche (agriculture, pêche, énergie) : ces deux catégories traduisent donc des éléments thématiques de la dépêche, mais selon des dimensions différentes. De façon pratique, dans les études précédentes qui ont utilisé ce corpus pour l'évaluation du clustering, les deux dimensions ont été utilisées : par exemple, (Lin et Cohen, 2010 ; Chen *et al.*, 2011) utilisent les catégories *industry*, alors que (Kim *et al.*, 2010 ; Tagarelli et Karypis, 2013) s'appuient plutôt sur les catégories *topic*. Nous avons, pour notre part, choisi d'utiliser une sous-partie des catégories *topic* et, plus précisément, les sous-catégories de la macro-catégorie GCAT (*Government/social*). Un alignement de ces catégories a été effectué manuellement avec les catégories de DMOZ et de Wikipédia, permettant de retenir au final 17 catégories.Les classes de référence ainsi obtenues sont alors effectivement des classes transverses aux diverses sources, mais la distribution des documents de chaque source à l'intérieur de chaque classe n'est alors pas homogène. L'évaluation peut donc être biaisée parce qu'un regroupement orienté par la source des documents peut donner de bons résultats pour les clusters pour lesquels une source est très majoritaire. Pour pallier ce biais, nous avons construit un corpus contenant des clusters de référence parfaitement équilibrés, en sélectionnant, pour chaque cluster, un nombre de documents de chaque source égal au nombre de documents de la source la moins fréquente. Ce corpus équilibré contient 38808 documents, 12936 par source, qui sont associés à 10 classes ; certaines classes qui ne contenaient pas assez de représentants d'un type de documents n'ont pas été gardées.

Pour vérifier la nature hétérogène de notre corpus, nous avons utilisé des mesures d'homogénéité et de similarité de corpus. Plus précisément, nous nous sommes appuyés sur les mesures proposées par (Kilgarriff, 2001) : nous présentons dans le tableau 2, les mesures obtenues sur ces collections avec les scores de corrélation de rang de Spearman des 500 mots les plus fréquents (à la différence de (Kilgarriff, 2001), nous ne considérons que les mots pleins, i.e. les mêmes mots qui sont utilisés pour la représentation vectorielle des documents). Les mesures d'homogénéité interne sont des moyennes sur les scores obtenus en comparant deux moitiés du corpus découpé aléatoirement, sur plusieurs découpages aléatoires différents, les mesures de similarités entre corpus sont obtenues de la même façon par des moyennes sur des découpages aléatoires (en prenant aléatoirement une moitié de chaque corpus).

On remarque bien dans ce tableau que les corpus issus d'une seule source sont très homogènes, alors que les distributions de vocabulaires sont très différentes d'un corpus à l'autre, avec des corrélations très faibles. Il est intéressant de noter que si la corrélation est quasiment nulle entre Reuters et les autres collections, elle est un peu plus élevée entre Wikipédia et DMOZ, qui sont deux collections issues du Web, mais elle reste très loin des mesures d'homogénéité intra-corpus.

	Reuters	Wikipedia	DMOZ
Reuters	0.968	0.103	-0.013
Wikipedia		0.940	0.247
DMOZ			0.952

Tableau 2. *Résultats des mesures d'homogénéité et de similarités entre les collections, établies par les scores de corrélation de Spearman.*

4.3. *Résultats du clustering*

4.3.1. *Sélection de traits dépendante du type des documents*

Nous présentons dans cette section les résultats des expériences menées sur la sélection de traits en utilisant l'information du type des documents, de façon à ne sélectionner que les traits les plus transverses aux différents types. Plus précisément, nous éliminons un nombre fixe k des termes en choisissant ceux qui ont un score de spécificité le plus élevé, en partant d'un clustering de référence s'appuyant sur une première sélection des termes selon leur fréquence en documents[4]. Les figures 1 et 2 présentent respectivement les scores de NMI et de *BCubed F-score* pour les différentes mesures de spécificité testées (présentées en section 3.2) et différentes valeurs de k.

On remarque sur ces figures une amélioration de la performance du clustering en supprimant des termes selon leur degré de spécificité selon les corpus. On remarque également qu'il y a un premier pic à 200 termes supprimés, avec, pour le score de spécificifé de l'information mutuelle, une amélioration du NMI de 0,394 à 0,413 et de 0,391 à 0,443 pour le BCubed-F-score. Ces deux figures illustrent une tendance générale vérifiée sur les deux mesures NMI et Bcubed F-score et montrent un comportement relativement proche pour les mesures du χ^2 (Chi-2), de l'information mutuelle (MI) et gain d'information (infoGain) en terme d'influence positive sur la qualité du clustering et de stabilité. Le comportement avec le score de fréquence en documents est plus chaotique, comme le montre la suppression de 200 et 500 termes sur le NMI. Le *log-likelihood ratio* (LLR) donne pour sa part de moins bons résultats. Il est également intéressant de noter une tendance générale à avoir un deuxième pic vers 700 documents, après une baisse des résultats, ce qui laisse penser que la présence ou l'absence de certains termes particuliers peut influer fortement la qualité des résultats : une analyse plus précise des termes retirés devrait être entreprise pour expliquer ce phénomène.

Le tableau 3 montre, pour illustration, les premiers mots supprimés (sur 200) selon la mesure du χ^2, en fonction du corpus où ils sont les plus discriminants. On voit dans ce tableau, d'une part, que la grande majorité (82%) des mots supprimés proviennent du corpus DMOZ et, d'autre part, que ces mots semblent effectivement très liés à la

4. Nous conservons à la base les termes ayant une fréquence en document entre 3 et 11000, cet intervalle ayant été fixé de façon empirique, en choisissant l'intervalle qui maximise le score NMI.

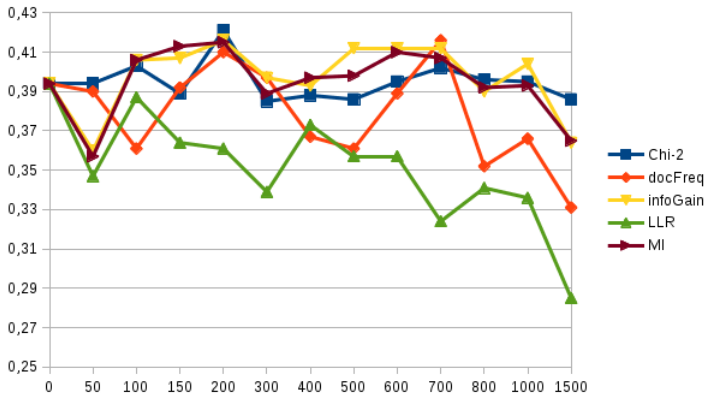

Figure 1. *Scores NMI, pour différentes mesures de spécificité et différents nombre de termes supprimés*

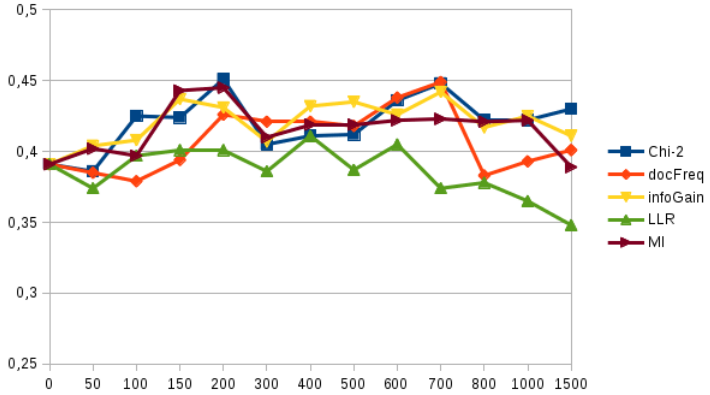

Figure 2. *Scores BCubed F-score, pour différentes mesures de spécificité et différents nombre de termes supprimés*

nature du corpus (documents web). Ce lien est moins évident pour les autres corpus mais il nous semble néanmoins clair qu'on ne supprime pas des mots fortement thématiques et que le thème premier des documents ne devrait pas être affecté par la suppressions de ces termes.

Les figures 3 et 4 présentent les distributions du nombre des documents provenant des différentes collections dans les clusters obtenus : la première figure pour le clustering de base, la seconde avec la sélection de traits utilisant la mesure du χ^2 pour supprimer 200 termes supplémentaires.

On constate dans cette comparaison qu'on obtient bien une meilleure homogénéité dans les clusters après la sélection des termes prenant en compte le type de

Source	Nb termes supprimés	Premiers termes supprimés (*stems*)
DMOZ	164	*site, privaci, copyright, search, email, click, page, informat, photo, video*
Reuters	30	*told, histori, percent, locat, newsroom, tuesday, reuter, school, list, thursday*
Wikipédia	6	*external, born, saturday, monday, debut, isbn*

Tableau 3. *Mots supprimés et associés au corpus où ils sont les plus discriminants selon le score du* χ^2

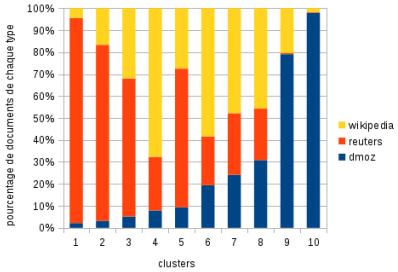

Figure 3. *Répartition des documents selon leur type dans les différents clusters pour un clustering de base*

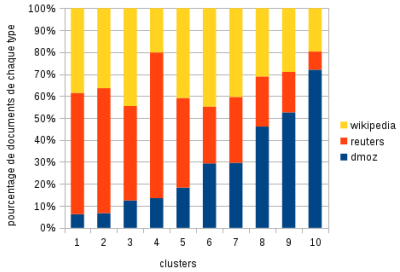

Figure 4. *Répartition des documents selon leur type dans les différents clusters, après suppression de 200 termes par la mesure* χ^2

documents : le nombre de clusters contenant des documents provenant d'une source particulière à plus de 60% passe de 7 clusters sur 10 avant filtrage à 2 clusters sur 10. De façon quantitative, cette amélioration peut être mesurée en calculant la statistique du test d'homogénéité du χ^2 pour mesurer l'écart des distributions par rapport à la distribution attendue (uniforme, par construction du corpus). Ces valeurs, données dans le tableau 4, confirment l'amélioration visible de la répartition des documents de différentes sources dans les clusters[5]. Ce résultat n'est pas forcément surprenant puisqu'on supprime effectivement les termes les plus spécifiques et qu'on limite donc la similarité des documents d'une même source. De façon générale, plus on supprime de termes, plus cette homogénéité augmente : il faut donc trouver un juste équilibre entre l'augmentation de l'homogénéité et la conservation d'un nombre suffisant de termes pour que la similarité soit pertinente.

5. Nous présentons les valeurs absolues de ces statistiques, sachant que, dans tous les cas, l'hypothèse d'une répartition uniforme des documents dans les clusters est rejetée.

	test d'homogénéité du χ^2
baseline	29444.30
chi2_200	9094.92

Tableau 4. *Valeurs des tests d'homogénéité du χ^2 sur la répartition des collections au sein des clusters*

	NMI	precision	rappel	F-score	BCubed F-score
baseline	0.394	0.646	0.245	0.356	0.392
chi2_200	0.421	0.576	0.247	0.346	0.451
MI+chi2+dfreq_100	0.416	0.657	0.271	0.383	0.430
MI+chi2+IG_100	0.422	0.657	0.267	0.380	0.432
IG+chi2+MI_200	**0.433**	0.619	**0.294**	**0.399**	**0.472**
IG+MI+chi2+dfreq+LLR_200	0.423	**0.666**	0.254	0.367	0.432

Tableau 5. *Résultat de clustering après fusion : NMI, F-score, BCubed F-score*

4.3.2. *Consensus Clustering*

Les mesures χ^2, gain d'information, *log-likelihood ratio*, information mutuelle et la fréquence en documents permettent de sélectionner des termes transverses aux collections et d'améliorer les résultats de clustering. Ces mesures étant différentes, les informations qu'elles fournissent sur les termes à supprimer peuvent être complémentaires. Nous évaluons ici cette complémentarité en opérant une fusion des résultats obtenus par une sélection de traits selon ces différentes mesures. Nous présentons dans le tableau 5 quelques combinaisons de regroupement pour lesquelles une amélioration de la qualité du clustering a été observée. La *baseline* correspond au clustering sans suppression de termes spécifiques aux collections. *chi2_200* fait référence au clustering en utilisant un filtrage de 200 termes par la mesure de spécificité du χ^2, qui est la solution de clustering de la section précédente qui présente globalement les meilleurs indices de qualité (dont NMI, F-score et bcubed F-score – toutes les mesures ne sont pas présentées par manque de place). Plusieurs combinaisons ont été testées, incluant trois ou cinq résultats de regroupements différents des données. Ainsi, *IG+MI+chi2+dfreq+LLR_200* désigne une combinaison de cinq résultats de clustering, dont le premier correspond au résultat pivot, ici produit avec un filtre par le score de Gain d'Information (IG), et dont les autres sont obtenus avec des filtres utilisant respectivement l'information mutuelle (MI), le χ^2, la fréquence documentaire et le *Log-Likelihood Ratio* (LLR).

Avec la méthode utilisée ici basée sur la fusion de clusterings, on n'obtient pas systématiquement des résultats supérieurs à ceux du clustering pivot. Les scores les plus élevés obtenus par les systèmes sont indiqués en gras.

	test d'homogénéité du χ^2
baseline	29444.30
chi2_200	9094.92
MI+chi2+dfreq_100	16394.19
MI+chi2+IG_100	16397.76
IG+chi2+MI_200	9804.89
IG+MI+chi2+dfreq+LLR_200	12201.74

Tableau 6. *Test d'indépendance des collections au sein des clusters*

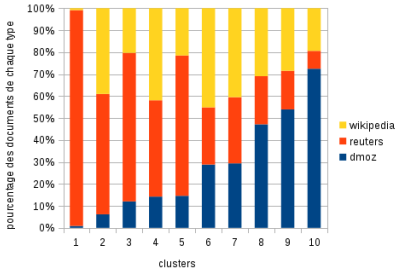

Figure 5. *Répartition des documents selon leur type dans les différents clusters pour le clustering IG+chi2+MI_200*

Ces résultats montrent qu' on peut obtenir un gain important de la qualité du clustering pour les systèmes combinant plusieurs résultats de clustering. Par exemple, pour le NMI, on va jusqu'à doubler le gain obtenu par la sélection des traits. On remarque également que l'accumulation de plus de résultats n'améliore pas forcément toujours la qualité, le consensus obtenu avec 5 résultats n'étant pas forcément meilleur que le meilleur obtenu avec 3 résultats. Par ailleurs, on constate que le gain de F-score observé par rapport à chi2_200 est surtout dû à une amélioration de la précision : avec le système de vote par majorité utilisé pour la fusion, les documents attribués à un même cluster par la majorité des clustering sont en effet plus probablement assignés au bon cluster que d'autres documents.

En ce qui concerne l'homogénéité observée dans les clusters, le tableau 6 présente les résultats de la statistique du test d'homogénéité du χ^2 sur les résultats obtenus par consensus, et la figure 5 présente la répartition des collections au sein des clusters pour le meilleur de ces résultats (IG+chi2+MI_200). On remarque que, si l'homogénéité des clusters est globalement plus importante que la *baseline*, elle reste inférieure à celle obtenue sans le *consensus clustering*. Cette diminution est en fait due, pour IG+chi2+MI_200 à un cluster particulier, composé à 98.26% de documents de Reuters : hormis ce cluster, l'homogénéité est comparable, pour des résultats de clustering bien meilleurs du point du vue des scores d'évaluation.

5. Conclusion

Nous nous intéressons dans cet article à une tâche de regroupement automatique de documents issus de collections hétérogènes. L'état de l'art montre que cette problématique a été peu étudiée. Nous proposons pour cette tâche une solution simple et indépendante de l'algorithme de clustering. Nous montrons par une évaluation que des mesures telles le χ^2, l'information mutuelle ou le gain d'information, qui sont habituellement utilisées pour déterminer des termes les plus spécifiques à des classes dans le cadre de la classification supervisée s'avèrent ici utiles. Ces mesures permettent de sélectionner des unités de représentation les plus transverses aux collections et favorisent le regroupement des données fondé sur autre chose que le genre ou la collection documentaire. Elles assurent une meilleure répartition des sources de documents au sein des clusters, tout en préservant la qualité globale du clustering. Le système de fusion des différents regroupements automatiques des documents peut produire des solutions de meilleure qualité.

Pour compléter cette étude, il faudrait comparer le gain apporté par notre approche à celui obtenu par des modifications directes de l'algorithme de clustering, sachant que, de façon générale, notre approche n'est pas incompatible mais plutôt complémentaire à ce type de méthode. Par ailleurs, cette approche s'appuie sur une connaissance *a priori* du type des documents. Une extension pourrait être d'appliquer cette méthode sur tout corpus, en effectuant par exemple un premier clustering qui déterminerait les types de documents, pour l'exploiter dans un second clustering : on se rapprocherait dans ce cadre d'une forme de clustering multi-facettes (*multi-view clustering*). Enfin, dans les collections hétérogènes, d'autres facteurs peuvent être source d'hétérogénéité comme, par exemple, la longueur des documents, s'il s'agit de regrouper des documents courts de type *tweets* avec d'autres documents plus développés. Ces facteurs ne sont pas pris en compte dans cette étude et demandent en général d'autres traitements spécifiques comme l'enrichissement automatique des documents. A nouveau, l'approche que nous proposons est complémentaire de ce type de méthode.

6. Bibliographie

Amigó E., Gonzalo J., Artiles J., Verdejo F., « A Comparison of Extrinsic Clustering Evaluation Metrics Based on Formal Constraints », *Inf. Retr.*, vol. 12, n° 4, p. 461-486, August, 2009.

Chen W.-Y., Song Y., Bai H., Lin C.-J., Chang E. Y., « Parallel Spectral Clustering in Distributed Systems », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, p. 568-586, 2011.

Dunning T., « Accurate Methods for the Statistics of Surprise and Coincidence », *COMPUTATIONAL LINGUISTICS*, vol. 19, n° 1, p. 61-74, 1993.

Forman G., « An Extensive Empirical Study of Feature Selection Metrics for Text Classification », *J. Mach. Learn. Res.*, vol. 3, p. 1289-1305, March, 2003.

Ghosh J., Acharya A., « Cluster Ensembles : Theory and Applications. », *Data Clustering : Algorithms and Applications*, Chapman & Hall, p. 551-570, 2013.

Ghosh R., Asur S., « Mining Information from Heterogeneous Sources : A Topic Modeling Approach », *Proceedings of the MDS Workshop at ACM SIGKDD*, 2013.

Hong L., Dom B., Gurumurthy S., Tsioutsiouliklis K., « A Time-dependent Topic Model for Multiple Text Streams », *Proceedings of ACM SIGKDD*, KDD '11, ACM, p. 832-840, 2011.

Katsavounidis I., Jay Kuo C.-C., Zhang Z., « A new initialization technique for generalized Lloyd iteration », *Signal Processing Letters, IEEE*, vol. 1, n° 10, p. 144-146, 1994.

Kilgarriff A., « Comparing Corpora », *International Journal of Corpus Linguistics*, vol. 6, n° 1, p. 1-37, 2001.

Kim Y.-M., Amini M.-R., Goutte C., Gallinari P., « Multi-view clustering of multilingual documents », *Proceedings of ACM SIGIR*, SIGIR '10, p. 821-822, 2010.

Kuhn H. W., « The Hungarian Method for the Assignment Problem », *Naval Research Logistics Quarterly*, vol. 2, n° 1–2, p. 83-97, March, 1955.

Lewis D. D., Yang Y., Rose T. G., Li F., « RCV1 : A New Benchmark Collection for Text Categorization Research », *J. Mach. Learn. Res.*, vol. 5, p. 361-397, December, 2004.

Lijffijt J., Papapetrou P., Puolamäki K., Mannila H., « Analyzing Word Frequencies in Large Text Corpora Using Inter-arrival Times and Bootstrapping. », *Proceedings of ECML PKDD 2011 (Part II)*, vol. 6912 of *LNCS*, Springer, p. 341-357, 2011.

Lin F., Cohen W. W., « A Very Fast Method for Clustering Big Text Datasets », *Proceedings of ECAI 2010*, p. 303-308, 2010.

Manning C. D., Raghavan P., Schütze H., *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.

Marx Z., Dagan I., Buhmann J. M., Shamir E., « Coupled Clustering : A Method for Detecting Structural Correspondence », *J. Mach. Learn. Res.*, vol. 3, p. 747-780, March, 2003a.

Marx Z., Dagan I., Shamir E., « Identifying Structure across Pre-partitioned Data », *Advances in Neural Information Processing Systems 16 (NIPS 2003)*, MIT Press, 2003b.

Slonim N., Tishby N., « Document Clustering Using Word Clusters via the Information Bottleneck Method », *Proceedings of ACM SIGIR*, SIGIR '00, p. 208-215, 2000.

Strehl A., Ghosh J., « Cluster Ensembles — a Knowledge Reuse Framework for Combining Multiple Partitions », *J. Mach. Learn. Res.*, vol. 3, p. 583-617, March, 2003.

Tagarelli A., Karypis G., « A segment-based approach to clustering multi-topic documents », *Knowledge and Information Systems*, vol. 34, n° 3, p. 563-595, 2013.

Taralova E., De la Torre F., Hebert M., « Source constrained clustering », *Computer Vision (ICCV), 2011 IEEE International Conference on*, IEEE, p. 1927-1934, 2011.

Vega-Pons S., Ruiz-Shulcloper J., « A Survey of Clustering Ensemble Algorithms », *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, n° 03, p. 337-372, 2011.

Yang Y., Pedersen J. O., « A Comparative Study on Feature Selection in Text Categorization », *Proceedings of ICML '97*, p. 412-420, 1997.

Zhai C., Velivelli A., Yu B., « A Cross-collection Mixture Model for Comparative Text Mining », *Proceedings of ACM SIGKDD*, KDD '04, p. 743-748, 2004.