Une nouvelle mesure pour l'évaluation des méthodes d'extraction de thématiques : la Vraisemblance Généralisée

Mohamed Dermouche*,** Julien Velcin* Sabine Loudcher* Leila Khouas**

*Laboratoire ERIC, Université Lumière Lyon2,
5 av. P. Mendès-France 69676 Bron Cedex, France
{julien.velcin, sabine.loudcher}@univ-lyon2.fr
**AMI Software R&D,
1475 av. A. Einstein 34000 Montpellier, France
{mde, lkh}@amisw.com

Résumé. Les méthodes dédiées à l'extraction automatique de thématiques sont issues de domaines variés : linguistique computationnelle, TAL, algèbre linéaire, statistique, etc. A ces méthodes spécifiques, peuvent s'ajouter des méthodes adaptées d'autres domaines, notamment de l'apprentissage automatique non supervisé. Les résultats produits par l'ensemble de ces méthodes prennent des formes hétérogènes : partitions de documents, distributions de probabilités sur les mots, matrices. Cela pose clairement un problème pour les comparer de manière uniforme. Dans cet article, nous proposons une nouvelle mesure de qualité, intitulée Vraisemblance Généralisée, pour permettre une évaluation et ainsi la comparaison de différentes méthodes d'extraction de thématiques. Les résultats, obtenus sur un corpus de documents Web autour des élections présidentielles françaises de 2012, ainsi que sur le corpus *Associated Press*, montrent la pertinence de la mesure proposée.

1 Introduction

Les documents textuels sont tellement abondants sur le Web que l'information pertinente est souvent difficile à retrouver. Dans l'objectif d'offrir une meilleure navigation dans les corpus de documents, que ce soit pour l'exploration du contenu ou la recherche d'information, l'extraction de thématiques (*topic extraction*) se distingue comme une tâche de fouille de textes dont l'objectif est d'extraire, automatiquement et sans catégories données *a priori*, des thématiques (sujets) à partir de grands corpus de documents. L'extraction de thématiques a été étudiée par différentes communautés, que ce soit celle de la fouille de données (Anaya-Sánchez et al., 2008), du Traitement Automatique des Langues (Blei et al., 2003), de la linguistique computationnelle (Ferret, 2006), de la recherche d'information (Zamir et al., 1997).

Les méthodes dédiées à l'extraction automatique de thématiques sont issues de plusieurs domaines : statistique, TAL, algèbre linéaire, etc. A ces méthodes spécifiques, peuvent s'ajouter des méthodes adaptées notamment de l'apprentissage automatique non supervisé. Les résultats produits par l'ensemble de ces méthodes prennent des formes hétérogènes : partitions, matrices, distribution de probabilités sur les mots, etc. Cela pose clairement un problème de

comparaison de ces résultats. Dans cet article, nous proposons une nouvelle mesure de qualité, la Vraisemblance Généralisée, qui permet d'évaluer et de comparer différentes méthodes d'extraction de thématiques. La mesure proposée est calculée dans un nouvel espace de description où les documents sont décrits par les thématiques et que nous appelons l'espace latent. Nous proposons également des opérateurs pour transformer les résultats des méthodes d'extraction de thématiques vers l'espace latent afin de calculer cette mesure.

La section 2 est consacrée à la présentation des principales méthodes d'extraction des thématiques. La section 3 présente les principales mesures de qualité ainsi que la nouvelle mesure intitulée Vraisemblance Généralisée. Les expérimentations et les résultats sont ensuite présentés en section 4. La conclusion et les perspectives de recherche sont données en section 5.

2 Méthodes d'extraction de thématiques

La plupart des méthodes d'extraction de thématiques nécessite que le corpus de documents soit mis sous forme d'une matrice V où les lignes représentent les documents et les colonnes représentent les mots (modèle vectoriel de (Salton et al., 1975)). Chaque élément V_{ij} de la matrice contient le poids du mot w_j dans le document d_i, qui reflète son importance dans le document. Le plus simple est de pondérer les mots par leurs fréquences d'apparition dans les documents (fréquence TF), même s'il existe d'autres types de pondération.

Il faut cependant noter que certaines méthodes d'extraction de thématiques, notamment celles issues de la linguistique computationnelle (Ferret, 2006), manipulent les documents sous forme d'un graphe où les nœuds représentent des unités linguistiques (mots, phrases, documents, etc.) et les arêtes représentent des relations entre elles, par exemple des relations sémantiques ou de co-occurrences. L'extraction des thématiques est ensuite effectuée en utilisant les algorithmes classiques issus de la théorie des graphes, comme le clustering spectral (Ng et al., 2002). Nous avons choisi de ne pas intégrer ce type de représentation dans notre travail, du moins pour le moment.

Dans cette présentation, nous proposons de regrouper les méthodes d'extraction de thématiques en trois grandes familles : les méthodes à base de distance, les méthodes à base de factorisation de matrices et les modèles de thématiques probabilistes.

2.1 Méthodes à base de distance

Les méthodes à base de distance se fondent sur le calcul d'une distance pour mesurer la similarité entre les documents. La plupart des méthodes de cette catégorie sont des méthodes de classification automatique non supervisée. Même si ce n'est pas leur vocation initiale, ces méthodes peuvent être utilisées pour l'extraction de thématiques en considérant que chaque classe définit une thématique et regroupe ainsi les documents qui y sont relatifs. La caractérisation des thématiques peut ensuite se faire en post-traitement en prenant par exemple les mots les plus fréquents dans chaque classe, ou en cherchant les mots les plus discriminants.

Dans les méthodes à base de distance, on trouve principalement les méthodes de partitionnement et les méthodes hiérarchiques. Les méthodes de paritionnement, comme l'algorithme des K-Means, commencent par répartir aléatoirement les documents sur un certain nombre de classes et, à chaque itération, les documents sont réaffectés de telle sorte que chacun soit dans la classe dont il est la plus proche (au sens de la mesure de similarité utilisée).

Plusieurs variantes des K-Means existent, comme FCM (*Fuzzy C-Means*), qui permet une classification floue des documents, c'est-à-dire un document n'est pas affecté à une seule classe mais il appartient à plusieurs classes avec différents degrés d'appartenance (Dunn, 1973). Les méthodes de partitionnement sont généralement de faible complexité, ce qui les rend adaptées aux grands volumes de données.

Les méthodes hiérarchiques procèdent à la construction des classes au fur et à mesure par agglomération, ou par division. En agglomération, chaque classe contient, au départ, un seul document. Les deux classes les plus proches, en termes de distance, sont ensuite fusionnées récursivement jusqu'à ce que tous les documents soient dans la même classe. En division, tous les documents sont dans une seule classe qui est divisée, récursivement, jusqu'à ce que chaque document soit dans une classe. Dans (Pons-Porrata et al., 2003), une méthode hiérarchique est proposée pour la classification de documents en se basant sur une distance qui prend en compte les entités temporelles et les noms de lieux. Les méthodes de classification hiérarchiques offrent la possibilité de contrôler la granularité des classes, et d'avoir ainsi des classes aussi fines ou grandes que souhaité. En revanche, les méthodes hiérarchiques souffrent du problème de complexité, ce qui les rend inadaptées aux grands volumes de documents.

Qu'elles soient à base de partitionnement ou hiérarchique, ces méthodes n'ont pas été initialement créées pour extraire des thématiques. Cependant, un simple post-traitement permet d'extraire des thématiques dans le sens où les centroïdes correspondent à des vecteurs dans l'espace du vocabulaire de mots. Cela explique la présence de ces méthodes dans cette étude.

2.2 Méthodes à base de factorisation de matrices

En algèbre linéaire, la factorisation de matrices est une approche qui peut permettre l'extraction des thématiques. Le principe général est de partir de la matrice d'occurences V, puis de trouver une factorisation de la matrice V en un produit de deux matrices W et H. H est une matrice dont les lignes sont constituées par des combinaisons de mots. Ce nouvel espace est appelé "espace sémantique latent" et il est défini par les thématiques, chaque thématique étant une combinaison de mots. W est une matrice de projection des documents dans le nouvel espace sémantique, où chaque élément W_{ij} représente le degré d'appartenance du document d_i à la thématique c_j.
L'analyse sémantique latente, LSA (*Latent Semantic Analysis*), permet de faire cette factorisation en effectuant une décomposition en valeurs singulières (Deerwester et al., 1990). Cependant, comme cette dernière peut produire des valeurs négatives, la méthode pose un problème d'interprétabilité des résultats (Lee et Seung, 1999). Pour contourner ce problème, la factorisation non négative de matrices, NMF (*Non-negative Matrix Factorization*), a été proposée par (Lee et Seung, 1999, 2001). NMF permet de trouver une factorisation non-unique d'une matrice non-négative V en un produit de deux matrices non-négatives W et H, de telle sorte que $V \approx WH$. L'objectif est de minimiser la fonction objectif J_{NMF} suivante :

$$J_{NMF} = \|V - WH\|^2 \tag{1}$$

(Lee et Seung, 2001) décrivent une méthode itérative basée sur NMF pour l'extraction de thématiques. Le problème de factorisation est ramené à un problème d'optimisation de la fonction J_{NMF} sous les contraintes de non négativité. Le problème est ensuite résolu en utilisant la méthode de Lagrange qui donne lieu aux deux règles de mise à jour suivantes :

$$W_{ij} \leftarrow W_{ij} \frac{(VH^T)_{ij}}{(WHH^T)_{ij}} \tag{2}$$

$$H_{ij} \leftarrow H_{ij} \frac{(W^TV)_{ij}}{(W^TWH)_{ij}} \tag{3}$$

2.3 Modèles de thématiques probabilistes

Les modèles de thématiques probabilistes (*probabilistic topic models*) sont une famille de modèles graphiques qui ont pour objectif la découverte de thématiques dans des corpus de documents. Le principe est de considérer un document comme un mélange probabiliste de thématiques latentes, c'est-à-dire un document est composé de plusieurs thématiques avec différentes proportions. Parallèlement à cela, chaque thématique est définie par une distribution de probabilités sur les mots. Par exemple, une thématique relative à la génétique associe des probabilités plus importantes sur les mots `ADN, gène, cellule`, etc. que sur les autres mots.
Un modèle de thématiques probabiliste peut être vu, d'un autre angle, comme un processus de génération de documents à partir d'un vocabulaire (ensemble fixe de mots, notés w_i), tout en prenant en compte le fait que chaque document est un mélange probabiliste de plusieurs thématiques, notées z_j. En supposant que les distributions $p(w_i|z_j)$ sont connues pour tout i, j, le processus simplifié de génération d'un document d est le suivant :

1. Se fixer une distribution de probabilités sur les thématiques $p(z_j|d)$
2. Pour chaque mot w à générer :
   (a) Choisir aléatoirement une thématique z parmi les z_j suivant la distribution fixée dans (1).
   (b) Choisir un mot w parmi w_i dans le vocabulaire suivant la distribution $p(w_i|z)$.

A partir de là, la procédure consiste à inverser le processus génératif en utilisant la loi de Bayes afin d'estimer les valeurs des paramètres $p(w_i/z_j)$ et $p(z_j/d)$. Ceci est réalisé en utilisant les techniques d'apprentissage et d'inférence des modèles graphiques probabilistes, comme le *Gibb's sampling* ou l'inférence variationnelle.

Les modèles de thématiques probabilistes proposés dans la littérature partagent globalement le principe génératif exposé ci-dessus, mais diffèrent principalement dans la manière de choisir les distributions de probabilités $p(z_i|d)$, $p(w_i|z_j)$. Dans PLSA (Hofmann, 1999), aucune hypothèse de la distribution des thématiques sur les documents n'est posée ; chaque document est traité à part. Dans LDA (Blei et al., 2003), chaque thématique est caractérisée par une distribution multinomiale sur les mots qui lui sont associés. LDA utilise une loi de Dirichlet pour permettre un choix judicieux des paramêtres des distributions multinomiales, et ainsi pallier les limites de PLSA.

Les modèles de thématiques probabilistes diffèrent également par leur structure. En effet, certains supposent l'existence d'autres variables latentes que les thématiques, par exemple des variables temporelles ou d'opinion, et permettent ainsi d'extraire ces connaissances en même temps que les thématiques.

Les trois familles de méthodes exposées ci-dessus ont des inspirations différentes. Néanmoins, il a été montré que des liens théoriques existent entre ces méthodes. La méthode NMF est équivalente à *Kernel K-Means*, une version des K-Means avec noyau (Ding et al., 2005). La méthode PLSA est équivalente à NMF en prenant la divergence de Kullback-Leibler dans la fonction objectif (Gaussier et Goutte, 2005).

3 Evaluation des méthodes d'extraction de thématiques

Les différentes méthodes d'extraction de thématiques produisent des résultats de forme hétérogène : partitions de documents, distributions de probabilités sur les mots, matrices, etc. Cela pose le problème de comparaison des résultats. Pour résoudre ce problème, nous proposons un nouvel espace de description commun aux différentes méthodes et une nouvelle mesure de qualité qui se calcule dans cet espace. Cela permet ainsi de comparer des approches de nature différente et de manière quantitative. Cette section présente les mesures de qualité existantes et la nouvelle mesure que nous proposons.

3.1 Mesures existantes

Les méthodes d'extraction de thématiques sont généralement évaluées de manière qualitative ou quantitative. L'approche qualitative a recours au jugement humain pour qualifier les thématiques sans donner aucun indice quantitatif pour comparer les méthodes entre elles. A contrario, l'approche quantitative permet de mesurer plus finement la qualité des modèles, qu'elle soit basée sur le jugement humain ou non. Le jugement humain est utilisé pour évaluer les thématiques selon deux critères : *word intrusion* et *topic intrusion* (Chang et al., 2009). Une des mesures quantitatives qui n'utilisent pas le jugement humain (automatiques) est la vraisemblance mais elle se calcule seulement sur les modèles probabilistes et sur les méthodes d'apprentissage de type EM (Dempster et al., 1977), ce qui n'est pas le cas de toutes les méthodes d'extraction de thématiques.

La problématique d'évaluation des thématiques se retrouve classiquement en apprentissage non supervisé avec les mesures recensées par (Halkidi et al., 2001). Celles-ci peuvent être réparties en deux catégories : mesures externes et mesures internes. Les mesures externes évaluent la qualité des résultats par rapport à une référence définie par les classes *a priori* des documents. Comme exemples de ce type de mesures on peut citer le F-score (moyenne harmonique du rappel et de la précision), l'entropie (mesure de désordre dans l'ensemble des thématiques) et la pureté (ratio moyen de la classe majoritaire dans chacune des thématiques). Les mesures internes ne font pas appel à des connaissances extérieures. Par exemple, l'inertie intra-classes est utilisée comme fonction objectif dans la méthode des K-Means, ou, la cohésion (Steinbach et al., 2000), qui mesure la similarité Cosinus entre les documents d'une même thématique.

Même si ces mesures peuvent être utilisées pour évaluer les thématiques, en considérant que chaque thématique correspond à une classe, elles ne sont pas dédiées à cette tâche. A notre connaissance, il n'existe pas à ce jour de mesure automatique qui permette d'évaluer toutes les méthodes présentées de manière uniforme, et ainsi de pouvoir les comparer.

3.2 Nouvelle mesure : la Vraisemblance Généralisée

La mesure que nous proposons, intitulée Vraisemblance Généralisée (VG), est une mesure quantitative interne qui permet d'évaluer plusieurs méthodes d'extraction de thématiques, même si ces dernières sont basées sur des modèles mathématiques différents. Le bien-fondé de la mesure VG repose sur le fait qu'il existe une analogie entre les différentes méthodes. En effet, toutes ces méthodes permettent, d'une manière ou d'une autre, de projeter les documents dans un espace de description formé par les thématiques et de décrire ces thématiques par des mots (cf. figure 1). Notre idée consiste à proposer la mesure VG qui se calcule dans l'espace latent ainsi que des transformations des résultats des méthodes vers l'espace latent.

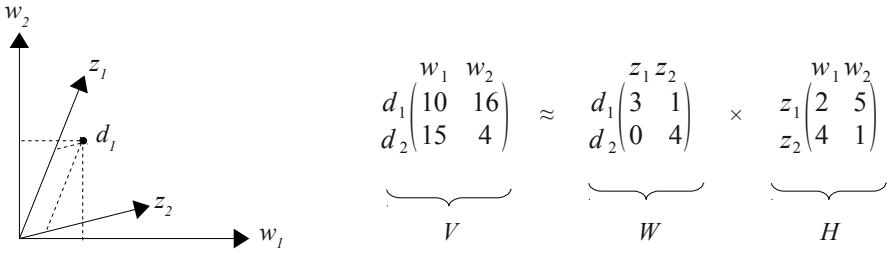

FIG. 1 – *Espace latent : les documents sont projetés dans l'espace latent caractérisé par les thématiques z_1, z_2 (matrice W) et les thématiques sont décrites par les mots (matrice H).*

La mesure VG est calculée à partir de deux matrices : la matrice de projection W, matrice de projection des documents dans l'espace latent et la matrice de l'espace latent H qui définit cet espace. La matrice H est caractérisée par un ensemble de vecteurs, pas nécessairement orthonormés, décrits dans l'espace des mots (cf. figure 1). Ce sont donc des vecteurs positifs ou nuls. La matrice W est caractérisée par un ensemble de vecteurs correspondant aux documents, décrits dans l'espace des thématiques. En d'autres termes, les matrices W et H sont telles que :

- W_{ik} est le score d'appartenance du document d_i à la thématique z_k.
- H_{kj} est le score d'appartenance du mot w_j à la thématique z_k.

Nous définissons trois transformations vers l'espace latent pour trois méthodes issues des principales approches présentées dans la section 2 (LDA, NMF et FCM). Pour la méthode LDA, W_{ik} est la probabilité $p(d_i|z_k)$, et H_{kj} est la probabilité $p(z_k|w_j)$. Pour la méthode NMF, les deux matrices sont directement obtenues par factorisation. Pour la méthode FCM, W_{ik} est le degré d'appartenance du document d_i à la classe z_k, et le vecteur H_{k*} est le centroïde de la classe z_k.

Les deux matrices W et H doivent être normalisées (si elles ne l'étaient pas déjà) afin d'avoir un même ordre de grandeur quelque soit la méthode utilisée et d'éviter ainsi un biais éventuel dans le calcul de la mesure :

- $\sum_{k=1}^{|Z|} W_{ik} = 1, \forall i \in \{1..|D|\}$ (normalisation des lignes de W).
- $\sum_{j=1}^{|W|} H_{kj} = 1, \forall k \in \{1..|Z|\}$ (normalisation des lignes de H).

Où Z est l'ensemble de thématiques et D est l'ensemble de documents. Sous ces hypothèses, nous définissons $score(d_i, w_j)$ le score de vraisemblance d'une occurrence du mot w_j dans le document d_i comme suit :

$$score(d_i, w_j) = \sum_{k=1}^{|Z|} W_{ik} \times H_{kj} \tag{4}$$

En d'autres termes, $score(d_i, w_j)$ est obtenu en multipliant la ligne de la matrice W qui correspond au document d_i ($i^{\text{ème}}$ ligne) par la colonne de la matrice H qui correspond au mot w_j ($j^{\text{ème}}$ colonne). Ensuite, le score de vraisemblance d'un document d est défini comme suit, V étant l'ensemble des mots du corpus (vocabulaire) :

$$score(d) = \prod_{w \in V} score(d, w)^{n(d,w)} \tag{5}$$

Où $n(d, w)$ est le nombre d'occurrences du mot w dans le document d. En passant au $\log$:

$$\log score(d) = \sum_{w \in V} n(d, w) \log score(d, w) \tag{6}$$

La mesure VG est basée sur la moyenne géométrique des scores individuels sur les documents, $score(d_i)$, chaque score étant lui-même un produit calculé sur chaque mot du vocabulaire (équation 5). La multiplication géométrique a donc la forme d'un produit de produits. Pour normaliser, il suffit de mettre à la puissance inverse du nombre de termes dans la multiplication. Celui-ci est égal à la double somme $\sum_{d \in D} \sum_{w \in V} n(d, w)$. Au final, la mesure VG est calculée avec la formule suivante :

$$VG(D) = \exp \left\{ \frac{\sum_{d \in D} \log score(d)}{\sum_{d \in D} \sum_{w \in V} n(d, w)} \right\} \tag{7}$$

où $\log score(d)$ est calculé avec la formule 6.

La mesure VG peut être calculée sur un corpus de test différent du corpus sur lequel les thématiques sont extraites (corpus d'apprentissage) mais ceci suppose que le modèle soit prédictif, c'est-à-dire capable d'affecter les nouveaux documents de test aux thématiques déjà extraites. Ceci n'est malheureusement pas le cas de toutes les méthodes, notamment les méthodes d'apprentissage non supervisé. Afin de mieux interpréter le résultat de la mesure VG, nous avons donc choisi, pour le moment, de l'évaluer en ne travaillant que sur le corpus d'apprentissage.

4 Expérimentations

Dans cette section, nous présentons le protocole expérimental (corpus, prétraitements, outils, paramètres des méthodes, etc.), ainsi que les résultats et la discussion.

Corpus	AP	Elections
Langue	Anglais	Français
Nombre de documents	2210	2777
Nombre de mots uniques	9794	9855

TAB. 1 – *Présentation des corpus AP et Elections.*

Thématiques	**immobilier**	**économie**	**vote des étrangers**	**sondages**
LDA	commun	economie	droit	sondage
	prix	crise	valeur	erreur
	immobilier	payer	vote	marge
	paris	argument	immigrés	institut
	politique	dire	étrangers	candidat
	crédit	marché	politique	journal
NMF	logement	euro	droit	sondage
	immobilier	politique	étrangers	institut
	construction	dollar	local	marge
	encadrement	part	municipal	erreur
	loyer	marché	gauche	harris
	prix	économie	non	ifop
FCM	spatial	page	fou	situation
	riom	divorce	marketing	copain
	municipal	seul	basculer	crever
	défaite	immatriculation	fier	croissance
	ajaccio	identifier	tronquer	sage
	bras	dollar	démontrer	attirer

TAB. 2 – *Exemple de thématiques découvertes par les trois méthodes sur le corpus Elections (nombre de thématiques = 50).*

4.1 Protocole expérimental

Les tests sont effectués sur deux corpus : AP et Elections. AP est un corpus de documents de l'agence de presse *Associated Press* (Harman, 1993), également utilisé dans (Blei et al., 2003). Elections est un corpus de documents Web (médias, blogs, réseaux sociaux, etc.), qui traitent des élections présidentielles françaises de 2012. Ces documents ont été collectés durant la période du 16/03/2012 au 16/04/2012 par la plateforme de veille AMIEI (http ://www.amisw.com). Le tableau 1 résume les contenus des deux corpus, après les prétraitements suivants :

- Suppression de mots outils (*stopwords*), par exemple `le, sur, dans`.
- Racinisation (*stemming*), par exemple les mots `logement, loger` deviennent `log`.
- Suppression des mots qui occurrent une seule fois dans le document.

Les tests sont réalisés en choisissant une méthode de chaque famille : LDA pour les modèles de thématiques probabilistes, NMF pour les méthodes à base de factorisation de matrices et FCM pour les méthodes à base de distance. Afin de limiter le risque de tomber dans des

optima locaux, le même test est réalisé 5 fois et la moyenne est retenue. Les paramètres de la méthodes LDA sont fixés comme suit : $\alpha = 50, \beta = 0.01$, nombre d'itérations $= 1000$. Les paramètres de la méthode FCM sont fixés comme suit : $m = 1.1$, nombre maximum d'itérations $= 20$. Pour exécuter LDA, nous nous sommes appuyés sur l'outil Mallet (McCallum, 2002). Pour NMF, nous avons utilisé notre propre implémentation, et pour FCM, nous avons utilisé le langage R (R, 2012).

Les deux types d'expérimentations réalisés sont les suivants :

– Test de comparaison : les trois méthodes FCM, NMF et LDA sont comparées suivant les scores obtenues par la mesure VG.
– Tests sur les cas extrêmes : deux cas extrêmes sont considérés : *Crisp* (chaque document appartient à une seule thématique à la fois), et *Uniforme* (chaque document appartient à toutes les thématiques avec le même score). Les résultats correspondant à ces deux cas extrêmes sont créés artificiellement en affectant le score 1 à la thématique qui maximise le score obtenu par NMF dans le cas *Crisp* et en affectant le score $\frac{1}{|Z|}$ à toutes les thématiques dans le cas *Uniforme*.

4.2 Résultats et discussion

Les résultats d'exécution des trois méthodes LDA, NMF et FCM sont représentés dans le tableau 2. La comparaison des trois méthodes par la mesure VG est représentée dans la figure 2. Les résultats sur les cas extrêmes sont représentées dans la figure 3.

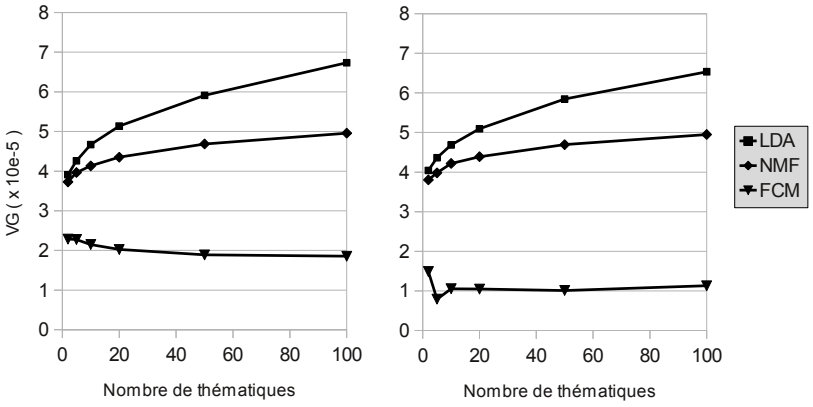

FIG. 2 – *Variation de la mesure VG en fonction du nombre de thématiques sur le corpus AP (à gauche) et Elections (à droite).*

Les méthodes LDA et NMF présentent un comportement similaire au vu de la variation de la mesure VG en fonction du nombre de thématiques (cf. figure 2). En effet, cette dernière augmente avec l'augmentation du nombre de thématiques. Ceci est en concordance avec l'intuition car un trop petit nombre de thématiques mène à mélanger plusieurs thématiques dans une seule, et donne ainsi des résultats de moins bonne qualité. En revanche, un grand

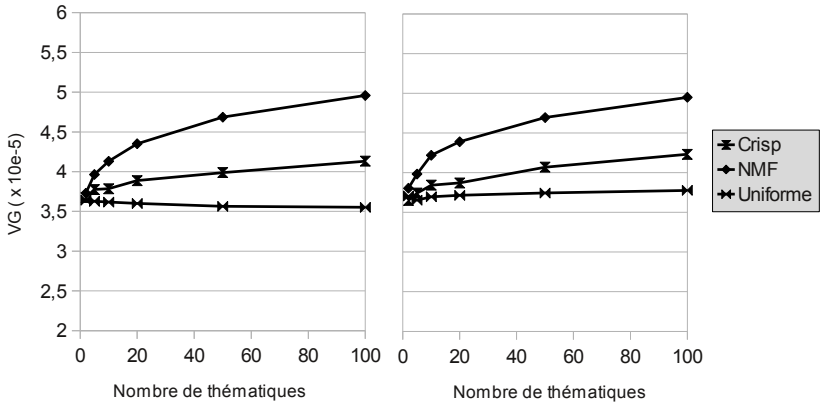

FIG. 3 – *Variation de la mesure VG en fonction du nombre de thématiques dans les cas extrêmes sur les corpus AP (à gauche) et Elections (à droite).*

nombre de thématiques permet de mieux séparer les thématiques, permet aux thématiques de petite taille d'émerger, et donne ainsi un résultat de meilleure qualité. Si le nombre de thématiques est encore plus grand (proche du nombre de documents), les résultats convergent vers un modèle où une thématique est extraite pour chaque document. La valeur de la mesure VG continue à augmenter sans pour autant que le résultat soit forcément de meilleure qualité. Ce problème est similaire au problème de surapprentissage (*overfitting*) connu dans le domaine de l'apprentissage statistique.

LDA demeure la méthode qui donne les meilleurs résultats, en termes de la mesure VG, par rapport à NMF et FCM, et ce sur les deux corpus (cf. figure 2). La qualité des résultats donnés par la méthode FCM est remarquablement inférieure, en termes de la mesure VG, à celle des deux autres méthodes. Ceci est conforme aux exemples donnés sur le tableau 2. En effet, les thématiques extraites par la méthode FCM sont mélangées et très difficiles à interpréter.

L'objectif du test sur les cas extrêmes est d'analyser le comportement de la mesure VG dans les deux cas extrêmes *Crisp* et *Uniforme* (cf. section 4.1). Suivant VG, *Crisp* et *Uniforme* sont des configurations moins bonnes que celle produite par NMF (cf. figure 3). Cela confirme que ces deux cas extrêmes ne donnent pas de bons résultats et qu'un bon ensemble de thématiques constitue en général un compromis entre les deux extrêmes, à savoir quelques thématiques pertinentes pour un document.

5 Conclusion

Les méthodes d'extraction de thématiques, étant issues de domaines variés, produisent des résultats de forme hétérogène, ce qui empêche leur comparaison de manière uniforme. Dans cet article, nous avons proposé une mesure d'évaluation, la Vraisemblance Généralisée, qui permet d'évaluer dans un cadre commun les méthodes d'extraction de thématiques. Pour calculer la

mesure, les résultats de ces méthodes sont transformés dans un espace latent qui plonge les documents dans l'espace latent des thématiques.

La mesure de qualité VG a permis de comparer trois méthodes d'extraction de thématiques (LDA, NMF et FCM) sur deux corpus différents. Les résultats ont donné l'avantage à la méthode LDA, suivie de NMF puis de FCM. Les résultats donnés par la méthode d'apprentissage FCM étaient d'une qualité inférieure, en termes de la mesure VG, par rapport aux deux autres méthodes. Ceci nous semble conforme avec une analyse qualitative des thématiques extraites par cette méthode. En effet, ces dernières étaient mélangées et très difficiles à interpréter.

Il serait intéressant, en complément à ce travail, de tester le comportement de la mesure VG sur des corpus de test (différents des corpus d'apprentissage). Cela nécessiterait la définition des opérations de prédiction pour les méthodes d'extraction de thématiques afin de pouvoir affecter les nouveaux documents aux thématiques. Il serait également intéressant d'envisager des analyses plus poussées afin de vérifier que l'on n'introduit pas de biais dans la comparaison des méthodes avec notre approche de transformation des résultats vers l'espace latent. En effet, il se peut que les transformations employées jouent en désavantage de certaines méthodes en dégradant ainsi artificiellement la qualité de leurs résultats.

Références

Anaya-Sánchez, H., A. Pons-Porrata, et R. Berlanga-Llavori (2008). A new document clustering algorithm for topic discovering and labeling. *Progress in Pattern Recognition, Image Analysis and Applications*, 161–168.

Blei, D., A. Ng, et M. Jordan (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research 3*, 993–1022.

Chang, J., S. Gerrish, C. Wang, et D. Blei (2009). Reading tea leaves : How humans interpret topic models. *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems 31*, 1–9.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, et R. Harshman (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science 41*, 391–407.

Dempster, A., N. Laird, et D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Ding, C., X. He, et H. Simon (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. *Proc. SIAM Data Mining Conf* (4), 606–610.

Dunn, J. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics 3*, 32–57.

Ferret, O. (2006). Approches endogène et exogène pour améliorer la segmentation thématique de documents. *Traitement Automatique des Langues 47*, 111–135.

Gaussier, E. et C. Goutte (2005). Relation between PLSA and NMF and implications. *Proceedings of the 28th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 601–602.

Halkidi, M., Y. Batistakis, et M. Vazirgiannis (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 107–145.

Harman, D. (1993). Overview of the first TREC conference. *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, 36–47.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57. ACM.

Lee, D. D. et H. S. Seung (1999). Learning the parts of objects by non-negative matrix factorization. *Nature 401*(6755), 788–91.

Lee, D. D. et H. S. Seung (2001). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems 13*(1), 556–562.

McCallum, A. K. (2002). MALLET : A Machine Learning for Language Toolkit.

Ng, A., M. Jordan, et Y. Weiss (2002). On spectral clustering : Analysis and an algorithm. *Advances in neural information processing systems 2*, 894–856.

Pons-Porrata, A., R. Berlanga-Llavori, et J. Ruiz-Shulcloper (2003). Building a hierarchy of events and topics for newspaper digital libraries. *Proceedings of the 25th European conference on IR research*, 588–596.

R, C. T. (2012). R : A Language and Environment for Statistical Computing.

Salton, G., A. Wong, et C. Yang (1975). A vector space model for automatic indexing. *Communications of the ACM 18*(11), 613–620.

Steinbach, M., G. Karypis, et V. Kumar (2000). A comparison of document clustering techniques. *KDD workshop on text mining*, 1–20.

Zamir, O., O. Etzioni, et O. Madani (1997). Fast and intuitive clustering of web documents. *KDD'97*, 287–290.

Summary

The topic extraction methods are issued from various domains: linguistics, linear algebra, NLP, statistics etc. In addition to these specific methods, we can adapt other methods, in particular from unsupervised machine learning. The results produced by all these methods are differently formed: document partitions, matrices, probability distributions over words. This difference causes a problem when trying to compare them uniformly. In this paper, we propose a new measure, named Generalised Likelihood, that allows evaluation and comparison of different topic extraction methods. The results obtained on both a corpus of Web documents about the french presidential election, 2012, and the *Associated Press* corpus show the relevance of the proposed measure.