Dimensionnalité intrinsèque dans les espaces de représentation des termes et des documents

Vincent Claveau

Dimensionnalité intrinsèque dans les espaces de représentation des termes et des documents

Vincent Claveau

IRISA-CNRS
Campus de Beaulieu, 35042 Rennes, France
vincent.claveau@irisa.fr

RÉSUMÉ. *L'examen des propriétés des espaces de représentation des documents ou des mots en RI (typiquement, $\mathbb{R}^n$ avec n très grand) fournit de précieuses indications pour aider la recherche. Récemment, plusieurs travaux ont montré qu'il était possible d'étudier la dimensionnalité réelle des données, appelée dimensionnalité intrinsèque, en certains points de ces espaces (Houle* et al.*, 2012a). Dans cet article, nous proposons de revisiter cette notion de dimension intrinsèque sous la forme d'un indice noté α dans le cas particulier de la RI et d'étudier son utilisation pratique en RI. Plus précisément, nous montrons comment son estimation à partir de similarités de type RI, peut être utilisée dans les espaces de représentations des documents et les espaces de représentations de mots (Mikolov* et al.*, 2013 ; Claveau* et al.*, 2014). Ainsi, nous montrons d'une part que l'indice α aide à caractériser les requêtes difficiles ; d'autre part, dans une tâche d'extension de requête, nous montrons comment cette notion de dimensionnalité intrinsèque appliquée à des mots permet de choisir au mieux les termes à étendre et leurs extensions.*

ABSTRACT. *Examining the properties of representation spaces for documents or words in IR (typically $\mathbb{R}^n$ with n large) brings precious insights to help the retrieval process. Recently, several authors have studied the real dimensionality of the datasets, called intrinsic dimensionality, in specific parts of these spaces (Houle* et al.*, 2012a). In this paper, we propose to revisit this notion through a coefficient called α in the specific case of IR and to study its use in IR tasks. More precisely, we show how to estimate α from IR similarities and to use it in representtion spaces used for documents and words (Mikolov* et al.*, 2013 ; Claveau* et al.*, 2014). Indeed, we prove that α may be used to characterize difficult queries; moreover we show that this intrinsic dimensionality notion, applied to words, can help to chosse terms to use for query expansion.*

MOTS-CLÉS : *Dimensionnalité intrinsèque, fonctions RSV, thésaurus distributionnels, extension de requête.*

KEYWORDS: *Intrinsic dimensionality, RSV scores, distributional thesauri, query expansion.*

1. Introduction

L'examen des propriétés des espaces de représentation des documents ou des mots en recherche d'information (typiquement, $\mathbb{R}^n$ avec n très grand) peut fournir de précieuses indications pour aider la recherche. Il est bien connu que la dimension de l'espace de représentation n'est pas représentative de celle des données ; dans un modèle vectoriel classique en RI, la dimension de l'espace de représentation est le nombre de mots différents dans la collection, mais il est souvent possible de représenter ces mêmes données dans un espace de plus petite dimension. C'est cette constatation qui est au cœur des approches de type LSI/LDA qui réduisent la dimension de l'espace initial (très creux) en un espace plus petit (et plus dense).

L'aspect qui nous intéresse dans cet article est la dimensionnalité intrinsèque des données, non pas du point de vue global de l'ensemble des données, mais localement sur une portion de l'espace. Pour cela, nous nous reposons sur les travaux récents de (Houle *et al.*, 2012a ; Amsaleg *et al.*, 2015) qui permettent de définir et d'estimer la dimensionnalité intrinsèque locale des données (section 2). Cette dernière dépendant de la façon dont le voisinage est défini, et donc de la métrique de distances utilisée, il est nécessaire de l'adapter aux particularités des fonctions de similarités RSV utilisée en RI (section 3). Nous montrons comment elle peut ensuite être utilisée pour analyser l'espace de représentation des documents en RI (section 4), par exemple pour repérer des requêtes difficiles. En section 5, nous menons le même travail sur les espaces de représentations des mots utilisés en sémantique distributionnelle (Claveau et Kijak, 2015). Enfin, nous montrons dans la section 6 une des nombreuses applications possibles de cette notion de dimension intrinsèque sur une tâche d'extension de requêtes à partir de lexiques distributionnels.

2. État de l'art

De nombreux travaux se sont intéressés à la caractérisation de la dimensionnalité intrinsèque d'ensemble de données dans des espaces de représentation de grande dimension. Ainsi, les méthodes de plongement (*embedding*) ou de projection construisent des espaces de plus petite dimension dans lesquels les données peuvent être projetés en respectant certaines contraintes de discriminabilité. C'est le cas des méthodes PCA/LSI/LDA (Deerwester *et al.*, 1990 ; Hoffman *et al.*, 2010) ou *manifold learning* (Scholkopf *et al.*, 1998 ; Roweis et Saul, 2000 ; Venna et Kaski, 2006). La dimension intrinsèque de l'ensemble du jeu de données est alors celle de ce sous-espace issu de la projection.

Récemment, (Houle *et al.*, 2012a) a proposé une mesure dite dimension d'expansion généralisée qui définit la dimensionnalité intrinsèque localement en examinant l'augmentation du nombre d'objets rencontrés autour d'un point de l'espace lorsqu'on augmente la distance à ce point. Pour illustrer cela, considérons deux boules de rayons ϵ_1 et ϵ_2 et centrées en x_1 et x_2 dans un espace $\mathbb{R}^m$, comme illustré en figure 1 (d'après (Houle *et al.*, 2012a)). Notons que si $\epsilon_2 = a \cdot \epsilon_1$ pour $a > 1$, alors le volume de

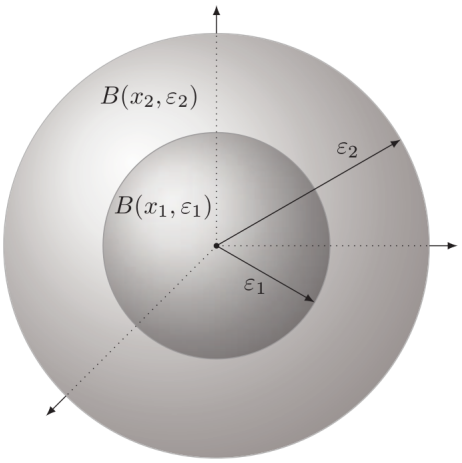

Figure 1. *D'après (Houle* et al.*, 2012a) ; deux volumes (voisinages de x_1 et x_2) dans un espace Euclidien à m dimensions, avec un centre commun ($x_1 = x_2$).*

la sphère extérieure est a^m fois plus grande que le volume de la sphère intérieure. Plus généralement, on peut s'intéresser au rapport entre ces volumes, qui s'exprime en fonction de la dimension m de l'espace :

$$\frac{volume(B(x, \epsilon_1))}{volume(B(x, \epsilon_2))} = \left(\frac{\epsilon_1}{\epsilon_2}\right)^m$$

De cela, on tire :

$$m = \frac{\ln(volume(B(x, \epsilon_1))) - \ln(volume(B(x, \epsilon_2)))}{\ln \epsilon_1 - \ln \epsilon_2}$$

L'idée au cœur de la dimensionnalité intrinsèque est de détourner ce calcul de la dimension m en considérant non pas le volume, mais le nombre de points contenus dans ce volume comme un estimateur. En notant $|B(x, \epsilon)|$ le nombre de points contenus dans le volume $B(x, \epsilon)$, on a donc :

$$\hat{m} = \frac{\ln |(B(x, \epsilon_1)| - \ln |B(x, \epsilon_2)|}{\ln \epsilon_1 - \ln \epsilon_2}$$

La dimension mesurée n'est plus celle de l'espace de représentation, mais celle propre aux données. Par ailleurs, il est important de noter que cette estimation est locale en un point x (en prenant le centre des boules $x_1 = x_2 = x$). Il ne s'agit donc pas d'une caractérisation de l'ensemble de points comme avec une PCA par exemple.

Ce modèle de dimensionnalité intrinsèque a été utilisé récemment dans des applications d'analyse et de construction de structures d'index pour des recherches par

similarités (Beygelzimer *et al.*, 2006 ; Houle *et al.*, 2012b ; Houle et Nett, 2013) et pour la détection d'anomalies (de Vries *et al.*, 2012).

3. Utilisation dans un contexte de RI

L'intérêt de la dimensionnalité intrinsèque pour la RI est sa capacité à caractériser le voisinage d'une requête en terme de documents l'environnant. Si l'espace autour de la requête est de très haute dimensionnalité, c'est le signe qu'une petite variation de distance change énormément l'ensemble des documents jugés proches de la requête. Autrement dit, une haute dimensionnalité implique une haute indiscriminabilité des documents autour de la requête (Houle *et al.*, 2012a). C'est cette propriété que nous voulons donc utiliser dans nos travaux, à condition de pouvoir adapter le cadre précédent aux cas particuliers des similarités utilisées en RI.

3.1. *Limites*

La définition de la dimensionnalité intrinsèque donnée précédemment vaut pour un espace dans laquelle la métrique utilisée est une distance, typiquement une distance L2 (euclidienne). C'est cette distance qui sert à définir les ensembles de points contenus dans les boules de différents diamètres. Par exemple, pour une boule centrée en x et de rayon $r > 0$, les points d_i considérés sont ceux à une distance L2 inférieure ou égale à r de x (soit encore de norme L2 inférieure ou égale à r si la boule est centrée sur l'origine).

En RI, la distance L2 est rarement utilisée comme fonction de pertinence (par la suite RSV, pour *Relevance Status Value*). En revanche, le cosinus a été largement utilisé comme fonction RSV dans le modèle vectoriel. Il permet de caractériser l'angle entre un vecteur requête et un vecteur document : deux vecteurs proches en terme d'angle (c'est-à-dire formant un angle petit) auront un cosinus élevé. Comme cela a été montré par (Houle *et al.*, 2012a), on peut adapter le calcul de dimensionnalité intrinsèque à ce cas des distances angulaires. L'idée est la même que précédemment : pour un vecteur requête donné, on examine le nombre de vecteurs dans ses voisinages à ϵ_1 et ϵ_2, c'est-à-dire avec une distance angulaire inférieure à un seuil ϵ_1 et ϵ_2.

Cependant, la plupart des fonctions RSV modernes peuvent s'écrire :

$$RSV(q,d) = \sum_{t \in q} w_q(t) \cdot w_d(t)$$

avec $w_d(t)$ et $w_q(t)$ qui sont respectivement les poids du terme t dans la requête q et dans le document d, comme illustré dans le tableau 1 (d'après (Lv et Zhai, 2011)).

	$w_q(t)$	$w_d(t)$
BM25+	$\frac{(k3+1)c(t,q)}{k3+c(t,q)}$	$\left(\frac{(k1+1)c(t,d)}{k1(1-b+b\cdot dl(d)/avdl)+c(t,d)}+\delta\right)\cdot\log\frac{N+1}{df(t)}$ avec $k1, k_3, b$ et δ des paramètres
PL2	$c(t,q)$	$\frac{tfn(t,d)\cdot\log_2(tfn(t,d)\cdot\lambda_t)+\log_2 e\cdot(1/\lambda_t-tfn(t,d))+0.5\log_2(2\pi\cdot tfn(t,d))}{tfn(t,d)+1}$ avec $tfn(t,d)=c(t,d)\cdot\log_2\left(1+c\cdot\frac{avdl}{dl(d)}\right)$ $c>0$ un paramètre de recherche et $\lambda_t=\frac{N}{c(t,C)}$
Dir	$c(t,q)$	$\log\left(\frac{\mu}{dl(d)+\mu}+\frac{c(t,d)}{(dl(d)+\mu)p(t\|C)}\right)$ $\mu>0$ un paramètre de lissage
Piv	$c(t,q)$	$\frac{1+log(1+log(c(t,d)))}{1-s+s\cdot dl(d)/avdl}\cdot\log\frac{N+1}{df(t)}$ si $c(t,d)>0$ et 0 sinon avec s un paramètre

Tableau 1. *Fonctions de pondération d'un terme dans une requête ou un document pour différents systèmes de RI de l'état-de-l'art : BM25+ (Robertson* et al.*, 1998 ; Lv et Zhai, 2011), Divergence From Randomness PL2 (Amati et Rijsbergen, 2002 ; Fang* et al.*, 2011), Modèle de langue avec lissage Dirichlet Dir (Zhai et Lafferty, 2001), Pivoted Normalization Piv (Singhal, 2001).*

avec les notations suivantes :

$c(t,d)$	nombre d'occurrences du terme t dans le document d
$c(t,q)$	nombre d'occurrences du terme t dans la requête q
N	nombre total de documents dans la collection
$df(t)$	nombre de documents contenant le terme t
$dl(d)$	longueur du document d
$avdl$	longueur moyenne des documents
$c(t,C)$	nombre d'occurrences du terme t dans la collection C
$p(t\|C)$	probabilité d'un terme t pour un modèle de langue de la collection

Ces fonctions de pertinence peuvent alors s'interpréter comme un simple produit scalaire entre le vecteur document d et le vecteur requête q, noté $\langle q,d\rangle$ par la suite. La seule différence par rapport au cosinus est donc de ne pas imposer une normalisation L2 des vecteurs. Cela peut se visualiser sur les figures 2 et 3 qui présente la distribution des normes L2 des documents de la collection Tipster (cf. section 4.1) et la rapport entre la longueur du document ($dl(d)$) et sa norme L2, le tout avec une pondération de type BM25+ (version modifiée de BM25 selon (Lv et Zhai, 2011)).

L'absence de normalisation a une conséquence importante puisqu'il apparaît difficile d'utiliser le même principe que précédemment pour calculer la dimension intrinsèque. En effet, pour une requête q donnée et deux valeurs de produit scalaire

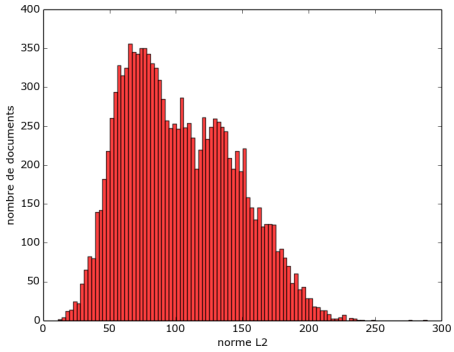

Figure 2. *Distribution des normes L2 des documents de la collection Tipster avec une pondération BM25+ (version modifiée selon (Lv et Zhai, 2011)).*

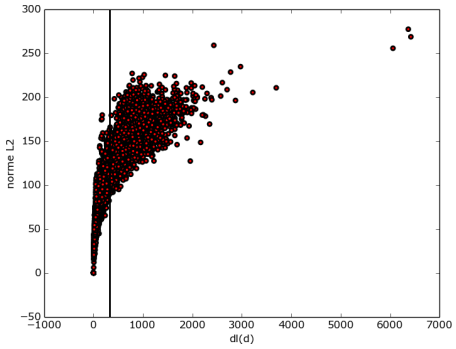

Figure 3. *Norme L2 des documents de la collection Tipster selon leur longueur $dl(d)$ avec une pondération BM25+ (version modifiée selon (Lv et Zhai, 2011)) ; la ligne verticale représente la longueur moyenne ($avgdl$).*

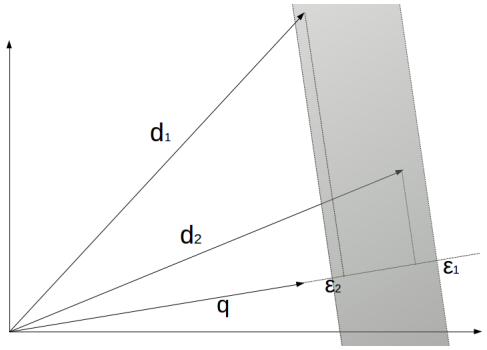

Figure 4. *En gris : portion de l'espace défini par l'ensemble des points dont le produit scalaire avec un vecteur q normé est entre ϵ_1 et ϵ_2.*

ϵ_1 et ϵ_2 ($\epsilon_1 \geq \epsilon_2$), la portion de l'espace pouvant contenir des points d_i tels que $\epsilon_1 \geq \langle d_i, q \rangle \geq \epsilon_2$ est infinie, comme illustrée en deux dimensions en figure 4.

3.2. *Estimation par l'exposant d'une loi de puissance*

Malgré la limite liée à l'utilisation de similarité sous forme de produits scalaires, on souhaite tout de même caractériser la dimension intrinsèque, ou du moins l'indis-

criminabilité, localement dans notre espace de représentation. Pour cela, nous nous inspirons des travaux de (Levina et Bickel, 2004 ; Amsaleg *et al.*, 2015) dans lesquels il a été montré qu'il était possible d'estimer la dimension intrinsèque à partir de la répartition des distances (par exemple L2 dans ces travaux) entre un vecteur requête et l'ensemble des autres vecteurs. Plutôt que de raisonner sur les rapports entre volumes, l'idée est de caractériser cette indiscriminabilité en examinant l'évolution du nombre de voisins (documents jugés proches de la requête) selon le score RSV. Cette courbe peut être interprétée (au sens des abscisses près) comme une fonction de répartition d'une variable aléatoire X représentant dans notre cas le score RSV entre le vecteur requête et un vecteur document. Plus précisément, puisqu'on s'intéresse au comportement local, on examine uniquement la portion de cette courbe RSV/nombre de voisins pour les plus grands scores RSV. L'hypothèse que nous faisons est que la distribution des RSV sur cette portion suit une loi de puissance, c'est-à-dire qu'elle s'exprime sous la forme :

$$f(x) = \lambda x^{-\alpha} \text{ avec } \lambda \text{ une constante et } \alpha > 1 \quad [1]$$

C'est cet exposant α qui est caractéristique de la dimensionnalité ou de l'indiscriminabilité des données.

Dans notre cas, nous disposons de n observations x_i, c'est-à-dire des valeurs de RSV entre une requête et ses n plus proches voisins (n meilleurs scores RSV). À partir de ces observations, il est possible d'estimer α. Parmi les différentes méthodes proposées dans la littérature, celle reposant sur la log-vraisemblance a été montrée comme la moins biaisée (Clauset *et al.*, 2009). Soient les observations x_i toutes supérieures à un seuil x_{min} on peut alors estimer α par :

$$\hat{\alpha} = 1 + n \cdot \left(\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}} \right) \quad [2]$$

Dans les expériences rapportées dans les sections suivantes, pour estimer α nous considérons $n = 100$ observations x_i, c'est-à-dire les 100 plus grands scores RSV.

4. Espace des documents

Dans cette section, nous souhaitons étudier l'utilisation de l'indice de dimensionnalité α, tel que défini dans la section précédente, dans un cadre classique de RI. Que les documents soient explicitement contenus dans des espaces euclidiens, comme avec le modèle vectoriel, ou que ce soit implicitement, au travers de fonctions de similarité pouvant s'interpréter comme des opérations dans des espaces euclidiens, nous montrons comment l'indice α peut-être utilisé pour caractériser les documents au voisinage d'une requête.

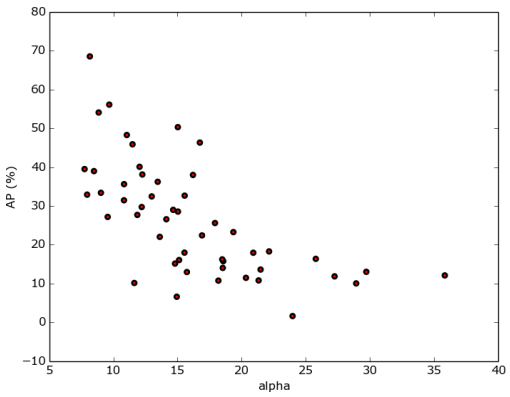

Figure 5. *Performances (AP) des requêtes selon leur indice α avec un modèle BM25+.*

coefficient	Valeur	p-valeur
Pearson r	-0.7150	5.43e-09
Spearman ρ	-0.7753	3.82e-11
Kendall τ	-0.5755	3.69e-09

Tableau 2. *Corrélations (et p-valeurs associées) entre AP et indice α.*

4.1. *Données et mesures d'évaluation*

La collection de RI que nous utilisons est celle développée pour le projet Tipster et utilisée dans le cadre de TREC. Elle contient plus de 170 000 documents et cinquante requêtes. Ces requêtes sont composées de plusieurs champs (la requête à proprement parler, un champ narratif détaillant les critères de pertinence) ; dans les expériences rapportées ci-dessous, nous n'utilisons que le champ requête. Les performances pour cette tâche de RI sont également classiquement mesurées en précision à différents seuils (P@x), R-prec, MAP.

4.2. *Indice de requête difficile*

La distribution des documents autour de la requête, ou plus précisément, la distribution des distances entre la requête et ses plus proches documents (selon le score RSV), peut aider à caractériser la difficulté de la requête. Pour vérifier cela, nous mettons en rapport l'indice α estimé (cf. équation 2) dans le voisinage de chaque requête avec l'Average Precision (AP) obtenue pour cette requête. Cela est illustré dans la figure 5 sur la collection Tipster, la fonction RSV utilisée est BM25+ (Lv et Zhai, 2011).

Le nuage de points obtenu illustre bien la dépendance attendue entre l'indice α, dérivé de la dimension intrinsèque, et la performance. Pour mesurer précisément cette relation, nous rapportons dans le tableau 2 les corrélations de Pearson, Spearman et Kendall (et les p-valeurs associées) entre la liste des requêtes ordonnées par AP et la même liste ordonnée par α. La corrélation inverse entre α et les performances obtenues pour la requête apparaît nettement : plus la dimension intrinsèque autour d'une

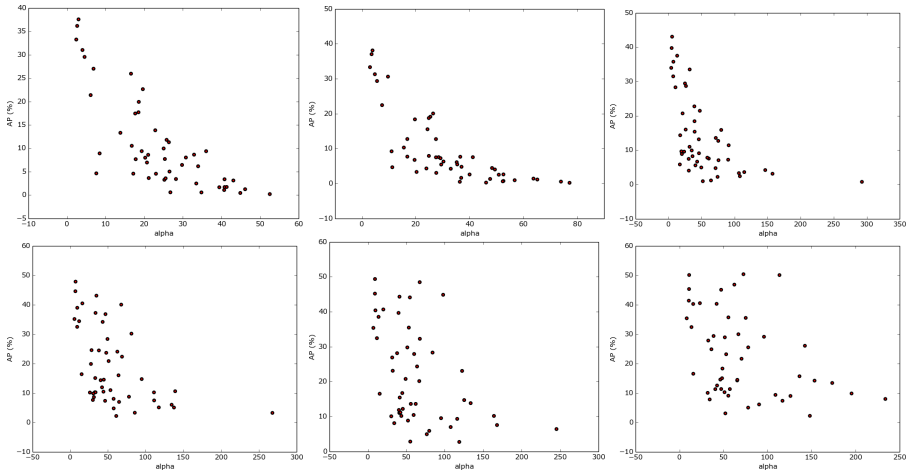

Figure 6. *Performances (AP) des requêtes selon leur indice α avec un modèle Dirichlet et dans le sens de lecture $\mu = 1, 25, 250, 1000, 2500, 4000$.*

requête est grande, moins bonne est la performance du moteur de recherche sur cette requête.

Les mêmes observations sont valables avec d'autres fonctions RSV. Dans les figures 6, nous montrons l'évolution α/AP en fonction du coefficient μ d'un modèle de langue avec lissage Dirichlet. Ces figures permettent d'observer un des effets du lissage : quand celui-ci est élevé, les documents ont tendance à avoir des valeurs similaires pour les termes de la requête et se trouvent donc tous à des distances comparables de la requête, ce qui a pour effet d'augmenter la dimensionnalité intrinsèque et donc l'indice α. Quand μ est faible, la corrélation (inverse) est donc très nette mais a tendance à disparaître quand μ augmente. Cela peut s'oberserver directement dans la figure 7.

5. Espace des mots

Depuis quelques années, de nombreux travaux se sont appliqués à projeter les mots dans des espaces de représentations continues, comme les espaces vectoriels. Il est ainsi possible de calculer, au sein de ces espaces, des proximités sémantiques à l'aide de distances/similarités vectorielles. Pour ce faire, ces techniques dites de plongement (*embedding*) reposent sur une hypothèse classique de sémantique distributionnelle : les mots proches partagent des contextes proches. Ainsi en examinant le contexte d'apparition des mots (c'est-à-dire les mots précédents et suivants), et en calculant des proximités entre ces contextes, ces méthodes d'analyse distributionnelle induisent des proximités entre les mots représentés par ces contextes.

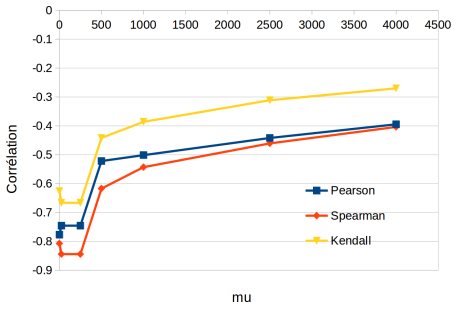

Figure 7. *Évolutions des valeurs de corrélations (r de Pearson, ρ de Spearman et τ de Kendall selon le coefficient de lissage μ dans le modèle de langue Dirichlet.*

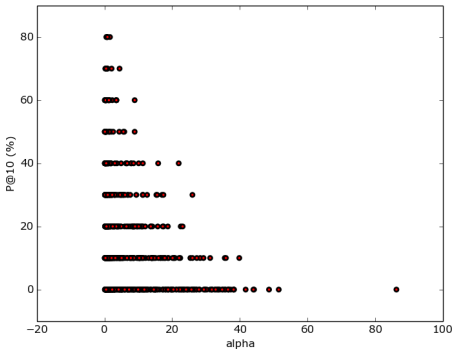

Figure 8. *Performances (P@10 sur WN+Moby) des mots-requêtes selon leur indice α avec un modèle BM25 selon la méthode de (Claveau et Kijak, 2015).*

Dans des travaux récents, (Claveau et Kijak, 2015) ont montré que les techniques de RI pouvaient être utilisées pour construire ces représentations de mots. En calculant par exemple des similarités de type Okapi-BM25 entre les ensembles de contextes de deux mots, ils ont montré qu'il était possible de construire des thésaurus distributionnels de qualité supérieure à l'état-de-l'art. C'est cette approche que nous reprenons ici. Comme nous l'avons fait pour les documents, nous nous intéressons dans cette partie à la dimension intrinsèque dans cet espace de représentation des mots. Puisqu'il repose lui-aussi sur des similarités RSV, nous reprenons la technique d'estimation de α pour caractériser localement l'indiscriminabilité en tout point de l'espace.

5.1. *Données et évaluation*

Les données que nous utilisons pour nos expériences de construction sont celles utilisées dans plusieurs travaux sur la sémantique distributionnelle. Cela va nous permettre de comparer nos résultats à ceux publiés. Le corpus utilisé pour collecter les contextes est le corpus AQUAINT-2 ; il est composé d'articles de presse en anglais et compte 380 millions de mots. Parmi eux, les mots que nous considérons pour entrées de notre lexique sont les noms communs apparaissant au moins 10 fois dans le corpus, soit 25 000 noms différents. Les contextes de toutes les occurrences de ces mots sont donc collectés ; dans les expériences rapportées ci-dessous, le contexte d'un mot est composé des deux mots à droite et deux mots à gauche du nom visé, en gardant leur position. Par exemple, dans l'extrait : "... all forms of restriction on freedom of expression, threats ...", les mots restriction-2, on-1, of+1, expression+2 sont ajoutés à l'ensemble des contextes freedom.

Comme dans les travaux de (Claveau et Kijak, 2015), nous utilisons conjointement WordNet (Miller, 1990) et Moby (Ward, 1996) pour l'évaluation intrinsèque des thésaurus produits. Ces deux ressources offrent des caractéristiques complémentaires : WordNet recense des liens sémantiques forts (synonymes ou quasi-synonymes) alors que Moby recense une plus grande variété de liens (hyperonymes, méronymes, co-hyponymie...). Une description détaillée des liens considérés par ces ressources est donnée par (Ferret, 2013) ou (Claveau *et al.*, 2014). Ainsi, WN propose en moyenne 3 voisins pour 10 473 des noms du corpus AQUAINT-2 et Moby 50 voisins en moyenne pour 9 216 noms. Combinées, ces deux ressources couvrent 12 243 noms du corpus avec 38 voisins en moyenne. Le nombre de noms dans les listes de référence et la variété des relations sémantiques considérées font de ces données un jeu d'évaluation très complet par rapport à d'autres *benchmarks* parfois utilisés tels que WordSim 353 (Gabrilovich et Markovitch, 2007).

Pour cette tâche de construction lexique sémantique, les mesures d'évaluation sont celles habituellement utilisées en RI.

5.2. *Indice de requête difficile*

Comme en section 4.2, nous examinons l'éventuel rapport entre l'indice α d'un mot et les performances obtenues en recherchant ses voisins sémantiques. Cela est illustré en figure 8 ; nous utilisons cette fois P@10 comme mesure de performance du fait de l'utilisation que nous faisons des voisins en section suivante. Le tableau 3 rapporte les valeurs de corrélation correspondantes.

coefficient	Valeur	p-valeur
Pearson r	-0.3607	4.46e-24
Spearman ρ	-0.4451	3.69e-37
Kendall τ	-0.3335	8.75e-42

Tableau 3. *Corrélations (et p-valeurs associées) entre P@10 et indice α*

coefficient	Valeur	p-valeur
Pearson r	-0.2268	4.65e-10
Spearman ρ	-0.8371	9.79e-195
Kendall τ	-0.6389	1.53e-148

Tableau 4. *Corrélations (et p-valeurs associées) entre nombre d'occurrences et indice α*

La corrélation négative entre α et P@10 est présente mais relativement faible. En toute généralité, l'indice α peut être utile mais n'est pas fiable à lui seul pour prédire la qualité du thésaurus distributionnel produit.

5.3. *Analyse fréquentielle*

Un autre élément d'analyse est l'étude du lien entre l'indice α et la fréquence du mots requête dans le corpus à partir duquel ses contextes sont collectés (et donc équivalent au $dl(d)$ dans cette tâche). En effet, plusieurs travaux précédents

(Ferret, 2013 ; Claveau *et al.*, 2014) ont montré l'influence de ce nombre d'occurrences sur la performance à la tâche de recherche des voisins sémantiques. Comme précédemment, nous reportons dans le tableau 4 les nombres d'occurrences des mots requêtes en fonction de leur indice α.

La corrélation est forte (et non linéaire puisque r est faible). Cela s'explique par le fait que le faible nombre d'occurrences, et donc le faible nombre de contextes collectés pour représenter un mot, se traduit par une description vectorielle insuffisamment détaillée du mot pour laquelle l'indiscriminabilité des voisins a donc tendance à être importante.

6. Extension de requête

Les lexiques sémantiques, tels que ceux générés par les techniques distributionnelles décrites précédemment, peuvent être utilisées dans beaucoup d'applications. En recherche d'information, ils peuvent notamment servir à étendre les requêtes : les voisins sémantiques des mots de la requête sont ajoutés à celle-ci. Depuis le travail séminal et les conclusions négatives de (Voorhees, 1994) dans lequel WordNet était utilisé, de nombreux travaux ont au contraire montré l'intérêt de l'extension de requête, que ce soit avec des lexiques construits manuellement ou générés automatiquement. Dernièrement, (Claveau et Kijak, 2015) ont montré que les lexiques distributionnels qu'ils généraient apportait des gains très substantiels en terme de MAP à la recherche, et que cette tâche de RI pouvait même être avantageusement utilisée comme une procédure d'évaluation des lexiques. Dans cette section, nous reprenons ce cadre applicatif en examinant comment notre indice α dérivé de la dimensionnalité intrinsèque peut être utilisé dans ce cadre.

6.1. *Mise-en-œuvre*

A des fins de comparaison, nous adoptons le même cadre expérimental que (Claveau et Kijak, 2015) :

– la collection de RI que nous utilisons est la même que précédemment, à savoir Tipster.

– le lexique distributionnel est celui étudié en section 5, identique à celui utilisé dans (Claveau et Kijak, 2015).

– le système de recherche d'information que nous utilisons est Indri (Metzler et Croft, 2004 ; Strohman *et al.*, 2005), connu pour offrir des performances état-de-l'art. Ce système probabiliste implémente une combinaison de modèle de langue (Ponte et Croft, 1998), tel que vu précédemment, et de réseaux d'inférence (Turtle et Croft, 1991) permettant d'utiliser des opérateurs tels que ET OU... Dans les expériences rapportées ci-dessous, nous l'utilisons avec des réglages standard, à savoir un lissage de Dirichlet ($\mu = 2500$). Ce système de RI offre l'avantage de disposer d'un langage de requête complexe qui nous permet d'inclure les mots du lexique distributionnel

Extension	MAP	R-Prec	P@5	P@10	P@50	P@100
Sans	21,78	30,93	92,80	89,40	79,60	70,48
avec extension	+13,80	+9,58	*+2,16*	+4,03	+5,58	+8,26
avec extension et filtrage 1	*+15,73*	*+9,27*	*+2,22*	*+4,96*	+9,63	+14,41
avec extension et filtrage 2	+22,83	+13,00	*+2,56*	+6,31	+14,10	+21,39

Tableau 5. *Gains relatifs de performance (%) par extension de requête sans filtrage et avec filtrage selon α sur les mots de la requête (filtrage 1) et sur les mots en extension (filtrage 2).*

en exploitant au mieux le modèle par réseau d'inférence à l'aide de l'opérateur dédié '`#syn`' qui permet d'agréger les comptes des mots considérés comme synonymes (voir la documentation d'Indri pour plus de détails). Pour supprimer les effets de flexions (pluriel) sur les résultats, les formes pluriel et singulier des noms de la requêtes sont ajoutées, que ce soit dans les requêtes non étendues avec les synonymes ou celles étendues par les voisins sémantiques.

– nous évaluons les performances avec les mesures classiques de RI, en comparant les résultats avec et sans extension.

6.2. *Expériences*

Beaucoup d'utilisations des indices α des requêtes elles-mêmes ou des mots de notre lexique distributionnel sont possibles. Nous rapportons ci-dessous les résultats de deux expériences où nous utilisons l'indice α des mots pour filtrer les extensions, selon deux mises-en-œuvre. Dans la première mise-en-œuvre, notée filtrage 1, nous calculons α pour chacun des mots de la requête originale, et nous ne proposerons des extensions (via le lexique distributionnel) que pour ceux inférieurs à un certain seuil (fixé au α moyen de tous les mots de toutes les requêtes dans l'expérience). Dans la seconde mise-en-œuvre, notée filtrage 2, le filtrage 1 est opéré, et en plus, nous filtrons également les mots ajoutés en extension selon leur indice α. Les résultats sont donnés dans le tableau 5 ; y sont rapportés les performances des requêtes orginales, et les gains par rapport à ces performances obtenues sans extensions. Des tests de significativité statistique (wilcoxon avec $p = 0.05$) ont été calculés : la version avec extension est comparée à la version sans extension, les versions avec filtrage (1 et 2) sont comparées avec la version avec extension (sans filtrage). Les résultats non significatifs sont en italique. De ces expériences, il ressort que les bons résultats de l'extension de requêtes par lexique distributionnel, déjà noté dans (Claveau et Kijak, 2015), est légèrement amélioré par un filtrage sur les mots à étendre (filtrage 1) mais sans que la différence soit statistiquement significative. En revanche, quand les extensions sont en plus elles-mêmes filtrées, le gain sur la version avec extension simple est alors plus important et statistiquement significatif. En pratique, un examen des requêtes étendues filtrées montre que les mots dont l'indice α dépasse le seuil sont effectivement ceux polysémiques ou généraux tels que : *choice, term, use, young*... Les bons résultats du

filtrage vont donc à l'encontre des faibles corrélations vues en 5.2 entre α et la précision sur les thésaurus de référence. Cette différence entre les résultats de l'évaluation intrinsèque (par thésaurus) et extrinsèque (via une tâche de RI dans notre cas) est en ligne avec les conclusions de (Claveau et Kijak, 2015).

7. Conclusion

Dans cet article, nous avons montré comment adapter la notion de dimensionnalité intrinsèque (Houle *et al.*, 2012a) aux similarités RSV utilisées en RI. En nous inspirant des travaux de (Amsaleg *et al.*, 2015), nous avons défini l'indice α que nous avons ensuite utilisé pour caractériser l'indiscriminabilité des voisins d'un point quelconque de l'espace de représentation. Appliqué à une tâche classique de RI (une requête pour trouver des documents), nous avons montré le lien entre cet indice calculé pour une requête et les performances du système de RI pour cette même requête. Nous avons appliqué la même approche en sémantique distributionnelle sur un espace contenant cette fois-ci des mots, mais utilisant les mêmes fonctions RSV (Claveau et Kijak, 2015). Nous avons alors montré l'intérêt de cet indice α pour améliorer les techniques d'extension de requête s'appuyant sur de tels lexiques distributionnels.

Beaucoup de perspectives sont ouvertes par ce travail. Tout d'abord, d'un point de vue théorique, la caractérisation de la dimensionnalité intrinsèque dans le cas de mesures de similarité (comme celles utilisées en RI) plutôt que de distances soulève plusieurs problèmes. Dans cet article, par analogie avec le travail de (Amsaleg *et al.*, 2015), nous avons utilisé l'indice α en supposant une distribution en loi de puissance. Même si cette distribution est vérifiée expérimentalement, le lien précis avec la dimensionnalité intrinsèque telle que définie par (Houle *et al.*, 2012a) et α est à préciser. D'autre part, on souhaite avoir une caractérisation locale de la dimensionnalité, mais l'estimation de α requiert un nombre minimal d'observations (RSV des plus proches voisins) pour être fiable. Ces deux contraintes sont parfois contradictoires pour certaines portions de l'espace dans lesquelles les plus proches voisins sont très éloignés.

D'un point de vue applicatif, on peut imaginer beaucoup d'utilisation de cette notion de dimensionnalité intrinsèque. Elle peut permettre de proposer des représentations de type LSI, LDA mais adaptées à la complexité locale. Plus pragmatiquement, lors d'une recherche en ligne, lors de la phase de rédaction de la requête, elle peut aussi aider à trouver les mots précis dans la requête qui provoque la plus forte montée d'α et demander à l'utilisateur d'utiliser un autre mot ou une autre formulation, ce qui permettra probablement de trouver des résultats plus pertinents, comme nous l'avons vu en section 4.

Remerciements

Nous tenons à remercier Laurent Amsaleg (IRISA-CNRS) et Teddy Furon (Inria Rennes) pour l'idée à l'origine de cet article et les discussions fructueuses tenues avec eux sur l'adaptation de leur approche aux métriques de RI.

8. Bibliographie

Amati G., Rijsbergen C. J. V., « Probabilistic models of information retrieval based on measuring the divergence from randomness », *ACM Trans. Inf. Syst.*, vol. 20, p. 357-389, October, 2002.

Amsaleg L., Oussama C., Furon T., Girard S., Houle M. E., Kawarabayashi K.-I., « Estimating Local Intrinsic Dimensionality », *21st Conf. on Knowledge Discovery and Data Mining, KDD2015*, Sidney, Australia, August, 2015.

Beygelzimer A., Kakade S., Langford J., « Cover trees for nearest neighbors », *Proc. of International Conference on Machine Learning (ICML)*, p. 97-104, 2006.

Clauset A., Shalizi C. R., Newman M. E. J., « Power-Law Distributions in Empirical Data », *SIAM Review*, vol. 51, n° 4, p. 661-703, 2009.

Claveau V., Kijak E., « Thésaurus distributionnels pour la recherche d'information et vice-versa », *Revue des Sciences et Technologies de l'Information - Série Document Numérique*, 2015.

Claveau V., Kijak E., Ferret O., « Improving distributional thesauri by exploring the graph of neighbors », *International Conference on Computational Linguistics, COLING 2014*, Dublin, Irlande, August, 2014.

de Vries T., Chawla S., Houle M. E., « Density-preserving projections for large-scale local anomaly detection », *Knowledge Information Systems*, vol. 32, n° 1, p. 25-52, 2012.

Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R., « Indexing by Latent Semantic Analysis », *Journal of the American Society for Information Science*, 1990.

Fang H., Tao T., Zhai C., « Diagnostic evaluation of information retrieval models », *ACM Trans. Inf. Syst.*, 2011.

Ferret O., « Identifying Bad Semantic Neighbors for Improving Distributional Thesauri », *51^{st} Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, Sofia, Bulgaria, p. 561-571, 2013.

Gabrilovich E., Markovitch S., « Computing semantic relatedness using wikipedia-based explicit semantic analysis », *20^{th} International Joint Conference on Artificial Intelligence (IJCAI 2007)*, p. 6-12, 2007.

Hoffman M., Bach F. R., Blei D. M., « Online Learning for Latent Dirichlet Allocation », *in* J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, A. Culotta (eds), *Advances in Neural Information Processing Systems 23*, Curran Associates, Inc., p. 856-864, 2010.

Houle M. E., Kashima H., Nett M., « Generalized expansion dimension », *Proc. of the 12th IEEE International Conference on Data Mining Workshops (ICDMW)*, p. 587–594, 2012a.

Houle M. E., Ma X., Nett M., Oria V., « Dimensional testing for multi-step similarity search », *Proc. of the 12th IEEE International Conference on Data Mining (ICDM)*, p. 299-308, 2012b.

Houle M. E., Nett M., « Rank cover trees for nearest neighbor search », *International Conference on Similarity Search and Appli- cations (SISAP)*, p. 16-29, 2013.

Levina E., Bickel P. J., « Maximum likelihood estimation of intrinsic dimension », *Advances in Neural Information Processing Systems (NIPS)*, 2004.

Lv Y., Zhai C., « Lower-bounding Term Frequency Normalization », *Proc. of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, ACM, New York, NY, USA, p. 7-16, 2011.

Metzler D., Croft W., « Combining the Language Model and Inference Network Approaches to Retrieval », *Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval*, vol. 40, n° 5, p. 735-750, 2004.

Mikolov T., Yih W.-t., Zweig G., « Linguistic Regularities in Continuous Space Word Representations », *2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL HLT 2013)*, Atlanta, Georgia, p. 746-751, 2013.

Miller G. A., « WordNet : An On-Line Lexical Database », *International Journal of Lexicography*, 1990.

Ponte J. M., Croft W. B., « A language modeling approach to information retrieval », *Proc. of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '98)*, p. 275-281, 1998.

Robertson S. E., Walker S., Hancock-Beaulieu M., « Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive », *Proc. of the 7th Text Retrieval Conference, TREC-7*, p. 199-210, 1998.

Roweis S. T., Saul L. K., « Nonlinear dimensionality reduction by locally linear embedding », *Science*, vol. 290, n° 5500, p. 2323-2326, 2000.

Scholkopf B., Smola A. J., Muller K.-R., « Nonlinear component analysis as a kernel eigenvalue problem », *Neural Computation*, vol. 10, n° 5, p. 1299-1319, 1998.

Singhal A., « Modern information retrieval : a brief overview », *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 2001.

Strohman T., Metzler D., Turtle H., Croft W., Indri : A language-model based search engine for complex queries (extended version), Technical report, CIIR, 2005.

Turtle H., Croft W., « Evaluation of an Inference Network-Based Retrieval Model », *ACM Transactions on Information System*, vol. 9, n° 3, p. 187-222, 1991.

Venna J., Kaski S., « Local multidimensional scaling », *Neural Networks*, 2006.

Voorhees E. M., « Query Expansion Using Lexical-semantic Relations », *Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, Springer-Verlag New York, Inc., New York, NY, USA, p. 61-69, 1994.

Ward G., « Moby Thesaurus », , Moby Project, 1996.

Zhai C., Lafferty J. D., « A study of smoothing methods for language models applied to ad hoc information retrieval », *Proc. of the SIGIR conference*, p. 334-342, 2001.