Méthodologie pour une représentation multi-dimensionnelle des documents

Benjamin Piwowarski

Méthodologie pour une représentation multi-dimensionnelle des documents

B. Piwowarski
benjamin@bpiwowar.net
LIP6/CNRS, Paris, France

La représentation des documents et questions en Recherche d'Information (RI) est restée une représentation majoritairement uni-dimensionnelle (i.e., vecteur). Cette représentation a des limites : Comment par exemple représenter un document qui traitent de plusieurs thèmes ou une question ambiguë ? Ces problèmes sont importants pour développer des systèmes de RI interactifs ou cherchant à diversifier les résultats. Les modèles actuels sont soit basés sur des heuristiques, soit sur des modèles latents tels que LDA (*Latent Dirichlet Allocation*) qui pré-supposent un nombre limités de thèmes pour décrire les documents. L'approche basée sur les probabilités dites "quantiques" permet d'établir des bases formelles pour une représentation multi-dimensionnelle des documents (ou plus généralement, des objets d'information) qui dépasse les limites évoquées plus haut. Cet article décrit la méthodologie QIA (*Quantum Information Access*) pour la représentation des documents, résume les résultats expérimentaux obtenus et décrit les perspectives.

1 Introduction

Les modèles de recherche d'information (RI) sont arrivés à maturité comme le montre le fait que les modèles probabilistes tels que BM25 [24] ou les modèles de langage [32] soient depuis plus d'une dizaine d'années utilisés comme base pour les comparaisons. Au niveau des modèles [1], les efforts portent aujourd'hui sur le développement de techniques permettant de prendre en compte la proximité entre les termes d'une question [3] ainsi que sur le développement de modèles permettant une interaction avec l'utilisateur [15].

Interagir avec l'utilisateur suppose deux mécanismes fondamentaux : (i) un modèle doit être capable de capturer la *diversité* thématique des questions et documents afin de présenter aux utilisateurs des résultats couvrant les différents thèmes correspondant au besoins d'informations possibles de l'utilisateur ; (ii) un modèle capable de prendre en compte les *interactions* entre l'utilisateur et le système pour permettre d'affiner les résultats au fur et à mesure. De nombreux modèles ont été proposés pour aborder les différents aspects liés à

1. Les efforts portant sur l'apprentissage sont pour nous orthogonaux à ces développements

l'interaction comme par exemple la mise à jour de modèles de langage [15], ou la prise en compte de retours négatifs [14]. De même, le problème de la diversité a été abordé à de nombreuses reprises, en utilisant deux grandes familles de techniques, celles cherchant à maximiser la dissimilarité entre les documents [1] ou celles basées sur une classification des documents dans différents groupes thématiques [13].

Malgré les nombreux travaux dans ces domaines, il n'existe pas encore de modèle établi qui permette de représenter de façon uniforme les documents et les besoins d'information pour gérer diversité et interaction. Une approche serait d'utiliser le formalisme probabiliste de la physique quantique comme proposé dans [22]. La physique quantique offre un formalisme qui lie géométrie (espaces de Hilbert) et probabilités (distribution de probabilité sur ces sous-espaces), tous deux sous-tendant la plupart des modèles de RI, à savoir les modèles vectoriels et probabilistes. Cela a amené van Rijsbergen [27] a proposer ce formalisme pour développer des extensions aux modèles de RI actuels. Conduits par cette motivation, la recherche en "RI quantique" cherche à utiliser le formalisme pour proposer des modèles qui permettent de résoudre certains problèmes de RI (diversité, interaction, multimédia).

Ces efforts ont conduit au développement de la méthodologie "Quantum Information Access" (QIA) pour représenter documents et besoins d'information. QIA répond au problème de la diversité et de l'interaction en exploitant la géométrie qui permet de représenter un document ou un besoin d'information comme un ensemble de vecteurs. Ces ensembles représentent la diversité du besoin d'information et des thèmes du documents ; la représentation du besoin d'information peut être mise à jour, permettant ainsi l'interaction. À notre connaissance, QIA est la seule méthodologie où les besoins d'information et les documents sont tous deux représentés de manière multi-dimensionnelle.

Cet article présente la méthodologie QIA et résume les résultats expérimentaux présentés dans [19, 21, 20, 18]. La méthodologie QIA peut être appliquée à des documents de différentes natures (texte, image, ou multimédia), bien que les expériences présentées dans les différents articles décrivant cette méthodologie, et donc rapportés ici, soient limitées au cas où le document est composé exclusivement de texte. Nous mettrons l'accent sur les hypothèses sous-tendant QIA et les conséquences des résultats expérimentaux déjà effectués.

Le plan de cet article est le suivant. Dans la section 2, nous présentons le contexte et les travaux liés à QIA. Dans la section 3, nous présentons succinctement le formalisme probabiliste "quantique" avant de décrire la méthodologie QIA. Enfin, nous présentons les principaux résultats expérimentaux dans la section 4 avant conclure par une discussion sur la méthodologie QIA dans la section 5.

2 Travaux liés

Le problème de la diversité a été abordée à de nombreuses reprises. Une approche typique est celle de [1] qui cherchent à maximiser la dissimilarité entre

les documents ou bien encore en en utilisant la théorie du portfolio [29] qui permet de minimiser le risque en diversifiant. Dans les deux cas, il faut une mesure de similarité entre deux documents. Une autre possibilité est de classifier les documents dans différents groupes thématiques [13] et de supposer que chaque groupe correspond à un aspect différent de la question. L'approche proposée par QIA fait partie de cette dernière catégorie, mais présente deux différences fondamentale : au lieu d'utiliser un nombre fini de thèmes, les documents *et* le besoin d'information sont représentés sous la forme d'un ensemble de vecteurs (voir section 3.2 pour une définition plus précise), et le nombre de thèmes n'est pas prédéfini. De plus, la diversité intrinsèque du besoin d'information ainsi que du document est préservée.

Au niveau de l'interaction, de nombreux modèles ont été proposés pour aborder les différents aspects liés à l'interaction comme par exemple la mise à jour de modèles de langage [15]. Globalement, il s'agit de mettre à jour un modèle vectoriel de l'utilisateur. Pourtant, des travaux tels que [9] ont montré que ce type de représentation n'est pas suffisante pour l'interaction, car il est difficile de prendre en compte les retours négatifs – cela peut se comprendre par le fait que s'il est possible, dans une certaine mesure, de représenter un besoin d'information sous la forme d'un vecteur, il est très difficile de faire de même pour représenter sa négation. Une solution proposée par [30] est de garder trace de tous les documents jugés non pertinents et d'utiliser une nouvelle mesure de similarité les prenant en compte. La méthodologie QIA traite le problème de manière plus élégante, en proposant qu'au départ le besoin d'information soit mal défini, et est "contenu" dans un sous-espace vectoriel de grande dimension. Au fur et à mesure des interactions, le sous-espace où peut se trouver le ou les besoins d'information diminue en taille : un jugement positif comme négatif est associé à un sous-espace qui permet d'éliminer les différentes possibilités qui ne correspondent pas aux interactions de l'utilisateur.

Tournons nous maintenant vers les travaux qui ont utilisé d'une manière ou une autre une représentation multidimensionnelle des documents.

La plupart des modèles en RI supposent qu'un document est mono-thématique. L'hypothèse mono-thématique se traduit par des facteurs de normalisation comme dans BM25 [24]. Cette normalisation n'est pas nécessaire dans le cas de QIA [20]. En effet, quand un document couvre plus de thèmes, il couvre plus de dimensions dans l'espace thématique. Par contre, s'il ne traite que d'un seul thème, il couvrira moins de thèmes. Une autre façon de voir cela est de considérer un document où du contenu est dupliqué. Dans le cadre de QIA, le sous-espace vectoriel correspondant à ce document sera exactement le même que dans le cas du même document sans contenu dupliqué.

Lorsqu'on veut capturer les différents thèmes traités par un document, il est possible de supposer que le document est une mixture de thèmes [31]. Toutefois, le nombre de thème est faible et fixe pour une collection. La méthodologie QIA repose sur une représentation multi-dimensionnelle où les thèmes peuvent être spécifiques, car ils correspondent chacun à un vecteur dans un espace de grande dimension ; et chaque document peut couvrir un nombre plus ou moins grand de dimensions en fonction du nombre de thèmes qu'il traite.

Une représentation plus explicitement multidimensionnelle a été développée par Che et al. [6] où le document est divisé en deux parties (aléatoirement), et une vue "stéréoscopique" du document est proposée. Pour calculer la similarité avec une question, les auteurs proposent de calculer de combiner les scores de pertinence des deux sous-documents. L'approche de QIA est là encore plus systématique et permet de ne plus avoir le côté arbitraire du nombre de "sous-documents" et de calcul du score.

Notre travail a des similarités avec les techniques où les documents sont représentés dans des espaces de dimensions réduites, telles que le *Latent Semantic Indexing* [8] ou *Latent Dirichlet Allocation* [4], car nous utilisons des techniques similaires (analyse spectrale) pour obtenir une approximation de la représentation d'un document ou d'un besoin d'information (section 3.2).

Plus proche de QIA, car exploitant les probabilités quantiques, Zuccon et al. [33] ont montré que l'hypothèse que des documents proches répondait plus souvent à une même question ("cluster hypothesis") tenait lorsque les documents étaient représentés comme des sous-espaces vectoriels. La principale différence avec QIA tient dans le principe utilisé pour construire ces sous-espaces. [16] a proposé de représenter un sous-espace vectoriel pour représenter un besoin d'information (ou plus exactement le contexte du besoin d'information). La probabilité qu'un document soit pertinent correspond à la probabilité que le vecteur représentant le document appartienne au sous-espace définissant le besoin d'information. La différence avec QIA est encore ici l'absence de méthodologie permettant de représenter les besoins d'informations, car le vecteur correspond à une représentation vectorielle classique de RI.

Finalement, en image, les sous-espaces sont déjà utilisés depuis longtemps pour certains problèmes de classification. Dans le travail de Belhumeur et al. [2], un visage est représenté par un sous-espace vectoriel qui est généré par plusieurs vues du même visage. La classification consiste a comparer la distance entre le vecteur représentant le visage à reconnaître et le sous espace vectoriel qui décrit le visage. La méthodologie QIA est relativement proche car nous calculons aussi des sous-espaces à partir de multiple "vues" sur le document, en exploitant une technique similaire mais plus systématique.

3 Méthodologie QIA

Dans cette section, nous présentons rapidement le formalisme probabiliste utilisé en physique quantique [2], avant de présenter et de discuter la méthodologie QIA.

3.1 Probabilités quantiques

La physique quantique décrit le comportement de la matière à une échelle (sub-)atomique en identifiant un état d'un système physique P à un sous-espace

2. Nous utiliserons le terme "probabilité quantique" qui est souvent utilisé pour ce formalisme

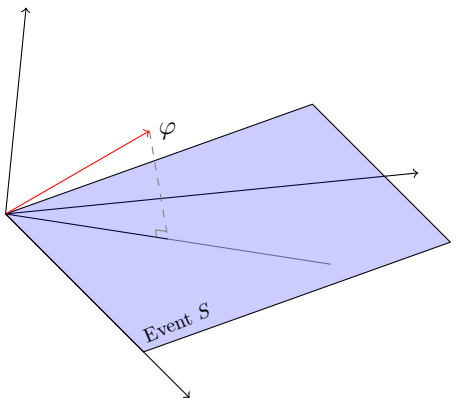

FIGURE 1: Probabilités quantiques : la projection de φ sur S

vectoriel dans un espace de Hilbert $\mathcal{H}$ – espace vectoriel défini sur le corps des nombres complexes. Ce sous-espace est très souvent représenté par un des vecteurs unitaires [3] φ qui génère ce sous-espace (il en existe une infinité) et nous suivrons cette convention : deux états φ_1 et φ_2 sont non compatibles (au sens probabiliste usuel) s'ils ne sont pas colinéaires. Un état définit de façon statistique les mesures qui peuvent être obtenues sur le système, comme par exemple la position d'une particule. Dans ce cas, le vecteur d'état φ associé avec cette particule détermine la probabilité qu'elle se trouve à un endroit donné.

Un événement est représenté comme un sous-espace S de l'espace de Hilbert $\mathcal{H}$. Si φ est totalement inclus dans le sous-espace, alors la probabilité de l'événement S est 1. Si φ est orthogonal au sous-espace, i.e. à n'importe quel vecteur de ce sous-espace, alors la probabilité est 0. De manière générale, la probabilité va être définie par la longueur de la projection de φ dans le sous-espace : plus le vecteur est orthogonal, plus la probabilité est faible. De façon formelle, la probabilité est définie par

$$q\left(S|P\right) = \left\|\widehat{S}\varphi\right\|^2 \tag{1}$$

où $\widehat{S}$ est le projecteur sur le sous-espace S, comme illustré dans la figure 1, et où le symbole q est utilisé pour distinguer la mesure de probabilité quantique q de la mesure de probabilité classique p. Il faut en effet noter que, même si l'état du système est connu et déterminé, i.e. nous connaissons le vecteur d'état φ qui caractérise le système, les événements ne sont pas certains. Ceci est une propriété du formalisme quantique.

Un système peut toutefois être dans un état non déterminé. Dans ce cas, il est possible de définir une distribution de probabilité $p\left(\varphi|P\right)$ sur l'ensemble des états que peut prendre le système P. La probabilité d'observer S est alors

3. ou plus exactement par la classe défini par l'équivalence $\varphi \sim \varphi' \equiv \exists\theta \in \mathbb{R}, \varphi = e^{i\theta}\varphi'$

donnée par

$$q(S|P) \quad = \quad \sum_{\varphi} p(\varphi|P) q\left(S|\varphi\right) = \sum_{\varphi} p(\varphi) \left\| \widehat{S}\varphi \right\|^2 \tag{2}$$

En suivant la terminologie classique, nous dirons que P définit une *densité de probabilité* (quantique). Notons que si le système est dans un état connu (le document ne contient qu'un seul aspect), alors l'équation ci-dessus se réduit à l'équation (1). Lorsque l'événement correspond à un seul état ψ (i.e., S est un sous-espace unidimensionnel), alors elle se réduit à (les vecteurs φ et ψ sont unitaires) :

$$q\left(S|P\right) = \left|\langle \psi, \varphi \rangle\right|^2 = \cos^2\left(\psi, \varphi\right) \tag{3}$$

qui est interprétée classiquement comme la probabilité de transition de l'état φ à l'état ψ.

La probabilité définie dans l'équation (2) est "quantique", elle n'obéit pas les lois de probabilités classiques. Ceci peut être vu simplement en montrant que la somme des probabilités de deux événements mutuellement exclusifs est supérieure à 1. Pour illustrer cela, considérons les trois événements associés avec les espaces uni-dimensionnels S_1 et S_2, associés aux vecteurs φ_1 et φ_2, de la figure 2. Si l'état du système est déterminé et égal à φ_1, alors $q\left(S_1\right) = 1$ et $q\left(S_2\right) = \left|\varphi_1 \cdot \varphi_2\right|^2 > 0$; la somme est bien supérieure à 1.

Du point de vue du calcul numérique, il est possible de montrer, en utilisant les propriétés de la trace tr et d'un projecteur $\widehat{S}$, que :

$$q\left(S|P\right) = \sum_{\varphi} p(\varphi|P) \mathrm{tr}\left(\varphi^\top \widehat{S} \varphi\right) = \mathrm{tr}\left(\widehat{S} \underbrace{\sum_{\varphi} p(\varphi|P) \varphi\varphi^\top}_{\rho_P} \right)$$

où φ^T est l'adjoint du vecteur φ et ρ_P est un opérateur linéaire qui définit de manière univoque la densité de probabilité de P. Cette équation montre qu'il est possible de séparer événement $\widehat{S}$ et densité de probabilité ρ_P. Elle montre également comment il est possible d'approximer ρ_P lors des calculs en utilisant une décomposition en valeurs propres :

$$\rho_P = \sum_i \lambda_i \varphi_i \varphi_i^\top \tag{4}$$

Cette approximation est utilisée en pratique pour estimer la représentation des objets d'information en prenant un négligeant les valeurs propres en dessous d'un certain seuil. Cette équation permet également de souligner un point important des probabilités quantiques : deux distributions $p(\varphi|P)$ peuvent avoir la même décomposition, et donc correspondre à la même densité de probabilité. Cela a des implications pour QIA qui seront détaillées dans la section suivante.

Finalement, il est possible de calculer une probabilité a posteriori, c'est à dire après avoir observé un événement S. D'un point de vue, cela correspond

à projeter les vecteurs φ correspondant à P dans le sous-espace vectoriel défini par S, en leur attribuant une probabilité proportionnelle à $q\left(S|\varphi\right)p\left(\varphi|P\right)$. Plus formellement,

$$q\left(\varphi|S;P\right)=\sum_{\psi/\varphi=\widehat{S}\psi/\left\|\widehat{S}\psi\right\|}q\left(S|\psi\right)p\left(\psi|P\right)/q\left(S|P\right) \tag{5}$$

qui se réduit à la formule de Bayes dans le cas particulier où tous les vecteurs φ tels que $p(\varphi|P)>0$ sont soit dans le sous-espace S, soit orthogonal à S. La formule de conditionalisation permet de définir comment prendre en compte l'interaction avec un utilisateur [22] en associant un sous-espace S à toute interaction entre l'utilisateur et le système. Nous de détaillerons pas cet aspect dans cet article.

3.2 La méthodologie QIA

Dans QIA, le concept de système ne fait pas référence à une entité physique, mais à un aspect d'un objet d'information. Un objet d'information représente toute agrégation de contenu : il peut s'agir d'un document comme d'un ensemble de documents ou de phrases. Nous considérons qu'un objet d'information est composé d'un ou plusieurs *aspects*. Un aspect peut faire référence à un thème (cas du texte), à la couleur ou la forme (cas d'une image), ou bien à des combinaisons comme par exemple couleur *et* forme.

Les aspects correspondent à des parties (ou fragments) de l'objet d'information qui "fait sens". Dans le cas du texte, un aspect (thème) correspond à un *"factoid"* [12], utilisé en résumé automatique ou dans les systèmes de question-réponse pour dénoter le montant d'information pertinente présente dans un résumé ou une réponse.

Ceci permet de définir les premières hypothèses de QIA :

Hypothèse 1. *Pour chaque type d'aspect, il existe un espace de Hilbert correspondant, comme par exemple l'espace thématique dénoté* $\mathcal{T}$ *ou encore l'espace couleur* et *forme* $\mathcal{C}\otimes\mathcal{F}$ *en utilisant un produit tensoriel*[4] *;*

Hypothèse 2. *La probabilité qu'un aspect* φ *soit similaire à un aspect* ψ *est donné par l'équation (3).*

La seconde hypothèse est familière en RI. Si le document est représenté par φ et la question par ψ, alors cela revient à dire que la probabilité que les thèmes soient similaires est donnée par le cosinus (au carré) entre φ et ψ. Ceci correspond au modèle vectoriel classique où le thème est un continuum qui va de "complètement hors-thème" (orthogonalité) à "exactement le même thème" (co-linéarité). La similarité doit donc être interprétée d'un point de vue intuitif comme la similarité donnée par le cosinus en RI et d'un point théorique

4. De façon intuitive, un vecteur φ de $\mathcal{C}\otimes\mathcal{F}$ correspond à un couple de deux vecteurs, $\varphi_C\in\mathcal{C}$ et $\varphi_F\in\mathcal{F}$ avec un produit scalaire égal au produit des produits scalaires dans les espaces $\mathcal{C}$ et $\mathcal{F}$.

comme la probabilité quantique. Notons également que contrairement au LSI, deux vecteurs φ et $-\varphi$ traitent exactement du même aspect dans QIA. Une discussion plus approfondie sur ce sujet peut être trouvée dans [34].

Nous supposons bien évidemment qu'un objet d'information est associé avec plusieurs aspects. Pour cela, nous formulons deux hypothèses supplémentaires :

Hypothèse 3. *Un objet d'information peut être décomposé en un ensemble de fragments. Ces fragments peuvent se chevaucher et être non connexes ; par exemple, un fragment peut être le premier paragraphe de chaque section, et un autre l'ensemble des paragraphes.*

Cela correspond à l'idée que la réponse à une question peut être n'importe quel partie cohérente d'un document [23].

Hypothèse 4. *Chaque fragment correspond à un ou plusieurs aspects qui ont plus ou moins d'importance, i.e. qui représentent plus ou moins l'objet d'information. Ceci est obtenu en associant chaque aspect à une probabilité.*

Il faut noter qu'une conséquence de l'équation (4) fait que deux distributions différentes peuvent correspondre à une même représentation "quantique". L'hypothèse ci-dessus est donc plus forte qu'il n'y parait.

Nous définissons maintenant les deux façon de considérer un objet d'information : soit comme une densité de probabilité, soit comme un événement.

3.2.1 Objet d'information comme densité de probabilité

En physique (quantique), les états sont exclusifs, i.e. un système peut être dans un seul état donné. De manière similaire, nous supposons qu'un objet d'information peut être associé un ensemble d'aspects, et que la probabilité que l'objet d'information traite d'un aspect donné est défini par la distribution de probabilité sur les aspects. Dans ce cas, un objet d'information O peut être vu comme une densité de probabilité quantique défini par la distribution sur les aspects $p\left(\varphi|O\right)$.

La probabilité qu'un objet d'information O traite d'un des aspects contenus dans S est défini par l'équation (2). Cette équation nous permet d'illustrer l'hypothèse fondamentale sur laquelle QIA repose. Considérons un cas simple où les événements sont des sous-espaces uni-dimensionnels S_{φ_i} définis par par un vecteur φ_i. Si un ou deux événements S_{φ_1}et S_{φ_2} ont une probabilité non-nulle, alors n'importe quel événement S_ψ, où ψ est défini par une combinaison linéaire de φ_1 et φ_2, aura une probabilité non nulle. Ceci est illustré par la figure 2. Dis d'une autre façon, un objet d'information qui traite de deux aspects φ_1 et φ_2 traitera avec une probabilité non nulle de n'importe quel aspect $\alpha_1\varphi_1 + \alpha_2\varphi_2$. Il est impossible de valider de manière théorique cette hypothèse, et seules les expériences permettent de déterminer si elle est valide.

3.2.2 Objet d'information comme événement

Il y a une autre manière de considérer un objet d'information dans QIA, à savoir comme un événement (quantique). Il est logique de supposer que le

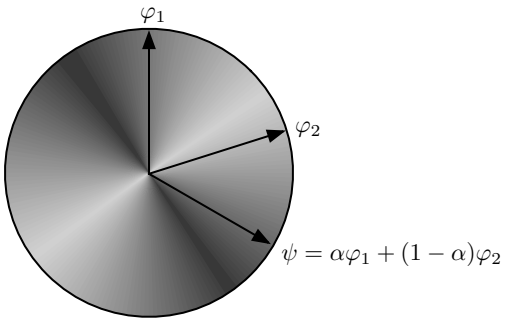

FIGURE 2: Densité (dimension 2) – les zones plus sombres dénotent une probabilité plus haute. Dans cette figure, on peut voir que la densité de probabilité change lentement en fonction de la "distance" avec la combinaison linéaire des deux vecteurs d'état.

sous-espace correspondant doit contenir tous les thèmes couverts par l'objet d'information, et pas plus. Plus formellement, nous voulons que $q\left(S_\varphi|O\right)$ soit égal à 1 pour tout aspect φ contenu dans l'objet d'information, et qu'il soit minimum pour tout aspect non directement contenu. Ceci amène à une solution unique, qui est que le sous-espace correspondant au objet d'information soit le sous-espace engendré par les aspects φ extraits de l'objet d'information. Nous appellerons S_O ce sous-espace.

Notons que lorsque deux aspects ψ et φ sont extraits de l'objet d'information O, cette construction fait que n'importe quel aspect qui est une combinaison linéaire de φ et ψ est considéré comme étant un thème de l'objet d'information. Cela constitue la dernière hypothèse faite par QIA :

Hypothèse 5. *Si deux aspects sont présents dans un objet d'information, alors n'importe quel combinaison linéaire de ces deux aspects est aussi présent dans le document.*

Finalement, cette construction est délicate car du bruit peut avoir un effet bien plus grand que dans le cas de la représentation sous forme de densité ; en effet, supposons que O ait deux fragments correspondant au même aspect, mais que le processus de d'extraction des thèmes renvoie deux vecteurs très légèrement différents, tels que φ et $\varphi + \epsilon$ où ϵ est négligeable. En toute rigueur, O sera représenté par un sous-espace vectoriel de dimension 2. Ceci peut être résolu en utilisant l'approximation donnée dans l'équation (4) en ignorant les valeurs propres les plus faibles.

4 Résultats

Dans cette section, nous rapportons les résultats obtenus en recherche d'information (ad-hoc) et en résumé automatique.

4.1 Recherche d'Information (ad-hoc et filtrage)

Les expériences sont décrites avec plus de détails dans [19, 20]. Nous décrivons ici la méthodologie expérimentale suivie et les principaux résultats obtenus. En résumé, nous considérons qu'un besoin d'information correspond à une distribution de probabilité sur l'espace thématique et qu'un document correspond à un événement (l'ensemble des thèmes abordés).

La représentation d'un document est obtenue de la façon suivante : (i) Les fragments sont des unités pré-définies comme les phrases, les paragraphes ou les sections (si disponible) ou bien encore comme des fenêtres glissantes de taille w avec un déplacement de o (mots 1 à w, o à $o+w$, $2o$ à $2o+w$, etc.) (ii) Chaque fragment correspond à un vecteur dans l'espace des termes (sac de mots) (iii) Le poids donné à chaque terme dans le fragment peut être binaire, tf, ou bien encore $tf \times \log \frac{N}{dtf}$ où tf est la fréquence du terme dans le fragment, dtf le nombre de document où apparaît le terme et N le nombre de documents. La pertinence d'un document est alors donnée par $p\,(D|Q)$ où Q correspond à la distribution de probabilité sur les thèmes composant la question (pour la tâche ad-hoc) ou les thèmes (pour le filtrage).

Dans le cas de la tâche ad-hoc, la représentation des questions est plus complexe, et correspond à l'idée suivante : l'ensemble des thèmes qui peuvent correspondre à un terme t d'une requête sont l'ensemble F_t des thèmes des fragments qui contiennent ce terme t. En supposant qu'il n'y a pas de raison qu'un thème soit plus important qu'un autre (distribution uniforme), nous avons

$$p(\varphi|Q_t) = \frac{1}{\text{card}\,(F_t)} \sum_{f \in F_t} \delta_{\varphi=\varphi_f}$$

où δ est égal à 1 si $\varphi = \varphi_f$.

Pour représenter une question avec plusieurs termes, il faut combiner les distributions de probabilité de l'équation ci-dessus pour plusieurs termes. Trois stratégies d'agrégation sont possibles, que nous illustrons ici dans le cas de deux termes t_1 et t_2 : 1. Mixture simple des probabilités ; 2. Mixture avec superposition : la probabilité d'un thème correspondant à la probabilité qu'il existe une combinaison linéaire d'un thème présent dans F_{t_1} et d'un présent dans F_{t_2}. Cela correspond à une généralisation de la combinaison linéaire utilisé pour agréger la représentation de deux textes en RI ; 3. Produit tensoriel : dans le cas spécifique de nos expériences, cela correspond à dire que la probabilité de pertinence d'un document est le produit des probabilités pour chaque terme (i.e. le document doit traiter des thèmes associés à t_1 et à t_2 pour être pertinent). Cette stratégie est celle qui est employée par exemple par les modèles de langages. Les expériences conduites sur TREC-1 à TREC-8 [20] montrèrent que la stratégie la plus robuste est d'utiliser un produit tensoriel, un découpage des documents en utilisant des fenêtres glissantes et un poids tf-idf pour les mots.

Les expériences sur la collection INEX [19] (où le produit tensoriel n'avait pas été utilisé) montrèrent que chaque type d'opérateur correspond à des types d'association différent : par exemple, la mixture avec superposition correspond

à des termes qui forment des concepts ("réseaux sociaux"), et la mixture correspond plus à questions où les termes correspondent à des aspects différents de la question ("tempête et dégâts"). Dans [5], nous avons essayé de systématiser cela en définissant une algèbre sur les différents opérateurs d'agrégation. L'idée était de pouvoir transformer automatiquement les requêtes de façon à exploiter les sémantiques différentes des trois opérateurs afin de les exploiter. Les résultats ont montré que s'il était potentiellement possible d'améliorer les résultats en fonction du type de question, il était difficile de le faire de façon automatique.

Finalement, dans [20], il est montré que le potentiel de la méthodologie QIA se trouve dans la formalisation élégante de l'interaction et des problèmes de diversité. S'il n'y a pour l'instant pas d'expériences qui ont été systématiques conduites dans ces domaines, des études préliminaires en diversité et en pseudo-retour utilisateur (*pseudo-relevance feedback)* ont donné de bons résultats. En filtrage, les résultats obtenus montrent que l'interaction (le filtrage peut être vu comme une interaction où le dernier document analysé est systématiquement jugé) peut être prise en compte de façon satisfaisante [21]. Dans cette tâche, la représentation de la question Q est tout simplement construit en construisant deux ensembles, l'un D^+ contenant les documents pertinents et l'autre D^- contenant les documents négatifs. Une densité de probabilité Q^+ est construite à partir de tous les fragments des documents de D^+. La distribution finale est obtenue par conditionalisation (équation 5) en observant que les thèmes ne font partie des documents D^- (et donc qu'ils font partie du sous-espace complémentaire D^-) :

$$p\left(\varphi|Q\right) = q\left(\varphi|\neg D^-; Q^+\right)$$

Les résultats montrent clairement que considérer l'ensemble D^- permet d'obtenir de bien meilleurs résultats, et donc que dans ce cas l'interaction peut être prise en compte par la méthodologie QIA de manière satisfaisante.

4.2 Résumé automatique

La tâche de résumé automatique avec laquelle nous avons fait des expériences a pour but d'extraire, à partir d'un ensemble de documents, l'ensemble des phrases qui résument le mieux les documents. Comme la plupart des travaux en résumé automatique extractif, nous considérons que les fragments de documents correspondent aux phrases des documents.

Pour un ensemble de phrases extraites R, nous pouvons définir l'ensemble des thèmes abordés par S_R comme un sous-espace (section 3.2). Nous considérons la distribution sur les thèmes définis par l'ensemble des documents à résumer $\mathcal{D}$.

Dans l'article [18], nous montrons qu'avec cette interprétation, les modèles basés sur une décomposition en valeur singulières proposées en résumé extractif, ont chacun des problèmes théoriques. Il est alors proposé un nouveau critère

$$R^* = \underset{R}{\operatorname{argmax}}\, p\left(R|\rho_{\mathcal{D}}\right)$$

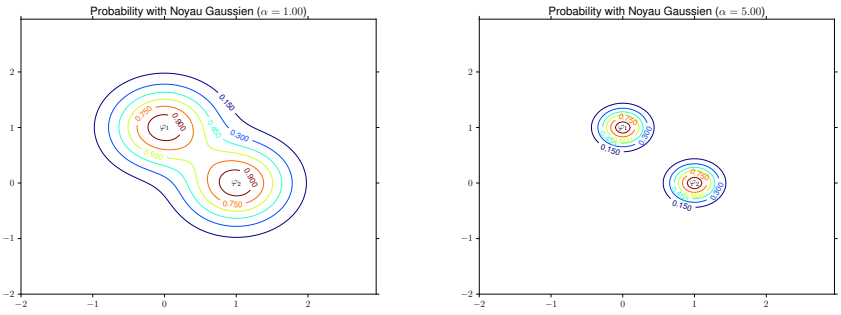

FIGURE 3: Probabilités obtenues avec un noyau gaussien (non normalisé) en fonction du paramètre α, i.e. $\langle \varphi_1, \varphi_2 \rangle_{\text{gaussien}} = e^{-\alpha \|x-y\|^2}$. L'événement est les sous-espace engendré par les vecteurs φ_1 et φ_2.

où R^* maximise la probabilités que les thèmes des documents $\mathcal{D}$ soient traités dans le résumé R^*. On peut voir que la probabilité est de 1 lorsque toutes les phrases sont sélectionnées, car le sous-espace engendré par l'ensemble des thèmes contient par définition tous les thèmes des documents. En pratique, il est impossible de maximiser $p\,(R|\rho_{\mathcal{D}})$ sur l'ensemble des résumés possibles, et un algorithme glouton est utilisé.

Un résultat obtenu dans cet article porte sur la représentation des documents qui n'avait pas été concluante dans les expériences de RI ad-hoc : la représentation basé sur un codage binaire/tf des mots dans un fragment ne fonctionne pas. La raison est simple. Utiliser un codage TF peut changer de manière conséquente la forme d'un sous-espace vectoriel. Considérons par exemple les pseudo-phrase , s_1 =“la phrase”, s_2 =“le paragraphe” and s_3 =“un paragraphe”. Avec un codage tf-idf, le sous-espace correspondant à $\{s_1, s_2\}$ serait très proche du sous-espace $\{s_1, s_3\}$ alors que cela ne serait pas le cas avec tf.

Les expériences ont montré que l'algorithme basé sur la méthodologie QIA obtient de meilleurs résultats que ceux obtenus lors des compétitions DUC 2006-08 par les systèmes participants, et que ceux obtenus par deux algorithmes performants, LexRank [10] et SNMF [28]. Ceci est particulièrement intéressant car pour la première fois il était possible à la fois d'analyser de manière fine des modèles existant de résumé et de les améliorer en utilisant les outils du formalisme quantique et de la méthodologie apportée par QIA.

5 Discussion

Dans cet article, nous sommes partis du constat qu'une méthodologie adéquate pour représenter documents et besoins d'information permettrait de combler un vide dans le domaine de l'accès à l'information : il n'existe en effet pas de modèle ou méthodologie bien établie qui permette d'aborder les tâches de recherche interactive et diverse.

Nous avons souligné le fait qu'une méthodologie possible est celle basée sur le formalisme des probabilités quantique et sur la méthodologie QIA (*Quantum Information Access*). Nous avons décrit le formalisme quantique, puis décrit cette méthodologie en mettant clairement en avant les hypothèses sur lesquelles il repose.

Nous avons discuté les résultats obtenus en RI (ad-hoc et filtrage) et en résumé automatique extractif. Pour la RI, les résultats montrent que la méthodologie permet d'obtenir des résultats équivalent à des modèles tels que BM25 [24]. Toutefois, la complexité numérique d'un modèle basé sur QIA fait que ce seul résultat n'est pas intéressant en soit, mais que ce sont plus les possibilités (diversité et interaction) qui rendent la méthodologie attractive.

Bien que cette piste de travail soit intéressante, il est apparu qu'un travail sur la représentation même du texte (le codage et l'utilisation de l'espace des termes comme espace thématique) est nécessaire pour s'assurer que la représentation est suffisamment fine pour pouvoir gérer ces différentes tâche de RI. Cette conclusion a été la conséquence du travail en résumé automatique [18] qui a montré l'importance de bien choisir l'espace des thèmes, et également du fait qu'en RI ad-hoc, les meilleurs résultats furent obtenus avec une fenêtre glissante alors qu'un découpage plus thématique serait souhaitable. Les résultats préliminaires obtenus en utilisant la transformation simple présentée dans [11] montre qu'une simple modification peut augmenter les résultats de façon significative.

Cela peut aussi se voir en considérant les différentes hypothèses sur lesquelles repose QIA. Les deux premières hypothèses sont assez classiques en RI, les deux suivantes paraissent intuitives (bien qu'il soit difficile de définir de façon précise qu'est-ce qu'un fragment qui fait "sens") mais c'est la dernière, l'hypothèse 5 qui est forte : "Si deux aspects sont présents dans un objet d'information, alors n'importe quel combinaison linéaire de ces deux aspects est aussi présent dans le document".

C'est afin d'explorer des espaces thématiques différents qu'une approche par noyau [26], qui permet de définir l'espace par le biais des produits scalaires et non plus de manière explicites (en calculant le vecteur), devient intéressante. . En effet, il est possible de redéfinir toutes les opérations présentées ici (calcul d'une probabilité, conditionalisation, décomposition en valeur propre) en utilisant les noyaux. Une librairie en C++ a été développée pour faciliter de tels calculs [17].

Jusqu'ici les noyaux ont été utilisés (explicitement ou implicitement) pour représenter des transformations permettant de capturer la sémantique [7] d'un texte, comme en les noyaux basés sur le LSI. Avec la méthodologie QIA, le fait que les documents et besoins d'informations puissent être représentés sous forme d'objet multi-dimensionnelle (i.e., plusieurs vecteurs) fait qu'il devient également intéressant de regarder les noyaux permettant de transformer la métrique associée à l'espace. Un exemple serait l'utilisation d'un noyau gaussien illustré par la figure 3 où événement (sous-espace) est généré par les deux vecteurs φ_1 et φ_2 (dimension 2). Alors qu'avec le produit scalaire classique, le sous-espace correspond à l'ensemble du plan, avec un noyau gaussien, il est possible de définir de manière plus fine ce qui est un thème "proche". Une autre possibilité serait d'utiliser un noyau de Fisher qui a montré son intérêt lorsqu'il s'agit de

combiner différents documents [25].

Finalement, pour explorer estimer les paramètres de ces noyaux (e.g., le α du noyau gaussien), il est possible d'utiliser des techniques simples d'apprentissage (descente de gradient), en utilisant les corpus standard de RI pour cela - maximiser la probabilité de pertinence des documents pertinents tout en minimisant la probabilité de pertinence des documents non pertinents.

Il sera également nécessaire d'exploiter différentes pistes (autres que la fenêtre glissante) pour définir les fragments d'un document. Une fois les meilleures représentations choisies, il sera possible d'explorer de façon plus systématique les tâches de diversité, d'interaction et de construction automatique de la densité de probabilité correspondant à une requête en suivant les idées exposées dans les articles [5, 20].

Références

[1] Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. Diversifying search results. In *WSDM* (New York, NY, USA, 2009), WSDM '09, ACM, pp. 5–14.

[2] Belhumeur, P., Hespanha, J., and Kriegman, D. Eigenfaces vs. Fisherfaces : recognition using class specific linear projection. *IEEE TPAMI 19*, 7 (1997).

[3] Blanco, R., and Boldi, P. Extending BM25 with multiple query operators. In *SIGIR '12 : Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (Aug. 2012), ACM Request Permissions.

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

[5] Caputo, A., Piwowarski, B., and Lalmas, M. A Query Algebra for Quantum Information Retrieval. In *Proceedings of the 2nd Italian Information Retrieval Workshop* (Jan. 2011).

[6] Che, L., Zen, J., and Tokud, N. A "stereo" document representation for textual information retrieval. *JASIST 5* (2006).

[7] Cristianini, N., Shawe-Taylor, J., and Lodhi, H. Latent Semantic Kernels. *Journal of Int Inf Sys 18*, 2 (2002), 127–152.

[8] Deerwester, S., Dumais, S., Furnas, G., and Landauer, T. Indexing by latent semantic analysis. *JASIST 41(6)* (1990).

[9] Dunlop, M. D. The effect of accessing nonmatching documents on relevance feedback. *ACM TOIS 15*, 2 (1997).

[10] Erkan, G., and Radev, D. R. Lexrank : Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research 22* (2004), 457–479.

[11] González, F. A., Caicedo, J. C., Amati, G., and Crestani, F. Quantum latent semantic analysis. In *Advances in Information Retrieval Theory - Proc. of ICTIR 2011* (2011), pp. 52–63.

[12] Halteren, H. V., and Teufel, S. Examining the consensus between human summaries : initial experiments with factoid analysis. In *HLT/NAACL-2003 Workshop on Automatic Summarization* (2003).

[13] He, J., Hollink, V., and de Vries, A. Combining implicit and explicit topic representations for result diversification. In *SIGIR* (New York, NY, USA, 2012), SIGIR '12, ACM, pp. 851–860.

[14] Karimzadehgan, M., and Zhai, C. Improving retrieval accuracy of difficult queries through generalizing negative document language models. In *CIKM'11* (Oct. 2011).

[15] Lv, Y., and Zhai, C. Adaptive relevance feedback in information retrieval. In *CIKM* (New York, New York, USA, 2009), ACM Press, p. 255.

[16] Melucci, M. A basis for information retrieval in context. *ACM TOIS 26*, 3 (2008).

[17] Piwowarski, B. The Kernel Quantum Probabilities (KQP) Library. *arXiv cs.MS* (Mar. 2012).

[18] Piwowarski, B., Amini, M.-R., and Lalmas, M. On using a quantum physics formalism for multidocument summarization. *JASIST 63*, 5 (2012), 865–888.

[19] Piwowarski, B., Frommholz, I., Lalmas, M., and van Rijsbergen, K. Exploring a multidimensional representation of documents and queries. In *RIAO* (2010).

[20] Piwowarski, B., Frommholz, I., Lalmas, M., and van Rijsbergen, K. What can Quantum Theory bring to IR ? In *CIKM* (2010), ACM.

[21] Piwowarski, B., Frommholz, I., Moshfeghi, Y., Lalmas, M., and van Rijsbergen, K. Filtering documents with subspaces. In *ECIR* (2010).

[22] Piwowarski, B., and Lalmas, M. A quantum-based model for interactive information retrieval. In *ICTIR* (2009).

[23] Piwowarski, B., Trotman, A., and Lalmas, M. Sound and complete relevance assessments for XML retrieval. *ACM TOIS 27*, 1 (2009).

[24] Robertson, S., and Zaragoza, H. The probabilistic relevance framework : BM25 and beyond. *Foundations and Trends in Information Retrieval 3*, 4 (2009).

[25] Seo, J., and Croft, W. B. Geometric representations for multiple documents. In *SIGIR* (2010).

[26] Smola, A. J., and Schölkopf, B. *Learning with Kernels*. MIT Press, 2002.

[27] van Rijsbergen, C. J. *The Geometry of Information Retrieval*. Cambridge University Press, 2004.

[28] Wang, D., Li, T., Zhu, S., and Ding, C. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of the* 31^{th} *annual international ACM SIGIR conference* (2008), pp. 307–314.

[29] Wang, J., and Zhu, J. Portfolio theory of information retrieval. In *SIGIR* (New York, New York, USA, 2009), ACM Press, p. 115.

[30] Wang, X., Fang, H., and Zhai, C. A study of methods for negative relevance feedback. In *SIGIR* (2008), ACM.

[31] Wei, X., and Croft, W. B. Lda-based document models for ad-hoc retrieval. In *SIGIR* (2006), ACM.

[32] Zhai, C. X. Statistical language models for information retrieval a critical review. *Foundations and Trends in Information Retrieval 2*, 3 (2008).

[33] Zuccon, G., Azzopardi, L., and van Rijsbergen, C. J. Semantic spaces : Measuring the distance between different subspaces. In *QI* (2009).

[34] Zuccon, G., Piwowarski, B., and Azzopardi, L. On the use of Complex Numbers in Quantum Models for Information Retrieval. In *ICTIR* (2011), pp. 346–350.