Méthodologie pour une représentation multi-dimensionnelle des documents

B. Piwowarski

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

RÉSUMÉ. *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 traite 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 qui pré-supposent un nombre limité 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.*

ABSTRACT. *The representation of documents and queries in Information Retrieval (IR) is mostly a one-dimensional (vector) representation. This representation has limitations: For example, how to represent a multi-topical document or an ambiguous query? These issues are crucial for the development of interactive systems or for diversification. Beside heuristics models that cannot be generalized, most representations are based on latent models such as LDA (Latent Dirichlet Allocation) that presuppose limited and small number of topics for documents, which implies a decrease in precision. The approach based on "quantum" probabilities can lay the groundwork for a formal multi-dimensional representation of documents (or more generally, information objects) that get over the limits mentioned above. The article discusses the QIA (Quantum Information Access) methodology for representing documents, summarizes the experimental results obtained so far and propose future directions.*

MOTS-CLÉS : *Accès à l'information, Représentation, Probabilités Quantiques*

KEYWORDS: *Information Access, Representation, Quantum Probabilities*

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 (Robertson *et al.,* 2009) ou les modèles de langage 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 ainsi que sur le développement de modèles permettant une interaction avec l'utilisateur (Lv *et al.,* 2009).

Interagir avec l'utilisateur suppose deux mécanismes fondamentaux : (i) 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) prise 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 l'interaction comme par exemple la mise à jour de modèles de langage (Lv *et al.,* 2009). 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 (Agrawal *et al.,* 2009) ou celles basées sur une classification des documents dans différents groupes thématiques (He *et al.,* 2012).

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 (van Rijsbergen, 2004). 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 résume les travaux sur la méthodologie QIA (Piwowarski *et al.,* 2010a, Piwowarski *et al.,* 2010c, Piwowarski *et al.,* 2010b, Piwowarski *et al.,* 2012). 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. Dans cet article, nous mettrons l'accent

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

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 succinctement le formalisme probabiliste "quantique" avant de décrire la méthodologie QIA et les hypothèses sur lesquelles elle repose dans la section 3. Nous présentons les principaux résultats expérimentaux dans les sections 4 et 5. Finalement, nous présentons une discussion sur la méthodologie QIA dans la section 6. Pour des raisons d'espace, la section sur les travaux liés (entre autres) a été écartée et peut être retrouvée dans la version longue de l'article (Piwowarski, 2013).

2. 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 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[2] 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 $q\left(S|P\right) = \left\|\widehat{S}\varphi\right\|^2$ où $\widehat{S}$ est le projecteur sur le sous-espace S 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 de plus être dans un état non déterminé et 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 donnée par

$$q(S|P) \quad = \quad \sum_{\varphi} p(\varphi|P) q\left(S|\varphi\right) = \sum_{\varphi} p(\varphi) \left\|\widehat{S}\varphi\right\|^2 \qquad [1]$$

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

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 $\left\|\widehat{S}\varphi\right\|^2$. 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) $|\langle\psi,\varphi\rangle|^2 = \cos^2(\psi,\varphi)$ qui est interprétée classiquement comme la probabilité de transition de l'état φ à l'état ψ.

La probabilité définie dans l'équation (1) 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. Si l'état du système est déterminé et égal à φ_1, alors $q(S_1) = 1$ et $q(S_2) = |\varphi_1 \cdot \varphi_2|^2 > 0$; la somme est bien supérieure à 1.

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 $q(S|\varphi)\,p(\varphi|P)$. Plus formellement,

$$q(\varphi|S;P) = \sum_{\psi/\varphi=\widehat{S}\psi/\left\|\widehat{S}\psi\right\|} q(S|\psi)\,p(\psi|P)\,/\,q(S|P) \qquad [2]$$

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 (Piwowarski *et al.,* 2009) en associant un sous-espace S à toute interaction entre l'utilisateur et le système. Finalement, le calcul numérique de telles probabilités repose sur des approximations (Piwowarski, 2012) basées sur les techniques spectrales d'algèbre linéaire. Nous de détaillerons pas ces aspects dans cet article.

3. 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".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*[3] *;*

Hypothèse 2. *La probabilité qu'un aspect φ soit similaire à un aspect ψ est donné par la formule* $\cos^2(\varphi, \psi)$

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 comme une 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 (Zuccon *et al.,* 2011).

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.

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 que 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.

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 définie 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 $p(\varphi|O)$.

3. 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 $\mathcal{F}$.

La probabilité qu'un objet d'information O traite d'un des aspects contenus dans S est défini par l'équation (1). 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.Dit 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.

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 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 quelle combinaison linéaire de ces deux aspects est aussi présente 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 lorsque les dimensions sont liés à des valeurs propres dont les valeurs les plus faibles peuvent être négligées.

4. Expérience en Recherche d'Information (ad-hoc et filtrage)

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 $p\left(D|Q\right)$ 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. 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 : (i) Mixture simple des probabilités ; (ii) 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 ; (iii) 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 probabilistes de RI. Les expériences conduites sur TREC-1 à TREC-8 (Piwowarski *et al.,* 2010b) 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 (Piwowarski *et al.,* 2010a) 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 (Caputo *et al.,* 2011), 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 (Piwowarski *et al.,* 2010b), 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. De plus, en filtrage qui peut être vu comme une interaction où le dernier document analysé est systématiquement jugé, les résultats obtenus montrent que l'interaction peut être prise en compte de façon satisfaisante (Piwowarski *et al.,* 2010c).

5. Résumé automatique

La tâche de résumé automatique que nous avons traitée 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). Nous considérons la distribution sur les thèmes définis par l'ensemble des documents à résumer $\mathcal{D}$.

Dans l'article (Piwowarski *et al.,* 2012), 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 $R^* = \text{argmax}_R\, p\left(R|\rho_{\mathcal{D}}\right)$ où R^* maximise la probabilité 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 $p\left(R|\rho_{\mathcal{D}}\right)$ 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 (e.g. LexRank (Erkan *et al.,* 2004)). 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.

6. Discussion

Dans cet article, nous avons résumé les travaux sur la méthodologie QIA (*Quantum Information Access*). 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 équivalents à des modèles tels que BM25 (Robertson *et al.,* 2009). 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 (Piwowarski *et al.,* 2012) 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 (González *et al.,* 2011) 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 (Smola *et al.,* 2002), 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 (Piwowarski, 2012).

Jusqu'ici les noyaux ont été utilisés (explicitement ou implicitement) pour représenter des transformations permettant de capturer la sémantique (Cristianini *et al.,* 2002) 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 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".

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 (Caputo *et al.,* 2011, Piwowarski *et al.,* 2010b).

7. Bibliographie

Agrawal R., Gollapudi S., Halverson A., Ieong S., « Diversifying search results », *WSDM*, WSDM '09, ACM, New York, NY, USA, p. 5-14, 2009.

Caputo A., Piwowarski B., Lalmas M., « A Query Algebra for Quantum Information Retrieval », *Proceedings of the 2nd Italian Information Retrieval Workshop*, January, 2011.

Cristianini N., Shawe-Taylor J., Lodhi H., « Latent Semantic Kernels », *Journal of Int Inf Sys*, vol. 18, n° 2, p. 127-152, 2002.

Erkan G., Radev D. R., « LexRank : Graph-based Centrality as Salience in Text Summarization », *Journal of Artificial Intelligence Research*, vol. 22, p. 457-479, 2004.

González F. A., Caicedo J. C., Amati G., Crestani F., « Quantum Latent Semantic Analysis », *Advances in Information Retrieval Theory - Proc. of ICTIR 2011*, p. 52-63, 2011.

He J., Hollink V., de Vries A., « Combining implicit and explicit topic representations for result diversification », *SIGIR*, SIGIR '12, ACM, New York, NY, USA, p. 851-860, 2012.

Lv Y., Zhai C., « Adaptive relevance feedback in information retrieval », *CIKM*, 2009.

Piwowarski B., « The Kernel Quantum Probabilities (KQP) Library », *arXiv*, March, 2012.

Piwowarski B., Méthodologie pour une représentation multi-dimensionnelle des documents (version étendue), Technical Report n° 00788414, HAL, 2013.

Piwowarski B., Amini M.-R., Lalmas M., « On using a quantum physics formalism for multi-document summarization », *JASIST*, vol. 63, n° 5, p. 865-888, 2012.

Piwowarski B., Frommholz I., Lalmas M., van Rijsbergen K., « Exploring a Multidimensional Representation of Documents and Queries », *RIAO*, 2010a.

Piwowarski B., Frommholz I., Lalmas M., van Rijsbergen K., « What can Quantum Theory bring to IR ? », *CIKM*, ACM, 2010b.

Piwowarski B., Frommholz I., Moshfeghi Y., Lalmas M., van Rijsbergen K., « Filtering documents with subspaces », *ECIR*, 2010c.

Piwowarski B., Lalmas M., « A Quantum-based Model for Interactive Information Retrieval », *ICTIR*, 2009.

Robertson S., Zaragoza H., « The Probabilistic Relevance Framework : BM25 and Beyond », *Foundations and Trends in Information Retrieval*, 2009.

Smola A. J., Schölkopf B., *Learning with Kernels*, MIT Press, 2002.

van Rijsbergen C. J., *The Geometry of Information Retrieval*, Cambridge UP, 2004.

Zuccon G., Piwowarski B., Azzopardi L., « On the use of Complex Numbers in Quantum Models for Information Retrieval. », *ICTIR*, p. 346-350, 2011.