Une approche d'extraction et de recherche d'information spatiale dans les documents textuels - Evaluation

C. Sallaberry*, M. Baziz*, J. Lesbegueries*, M Gaio*

** Laboratoire d'Informatique-Université de Pau (UPPA)*
Avenue du doyen Poplawski, BP 575
64013 PAU Cedex, France

{christian.sallaberry,mustapha.baziz,julien.lesbeguerie,mauro.gaio}@univ-pau.fr

RÉSUMÉ. *Ce papier propose une approche d'Extraction d'Information (EI) et de Recherche d'Information (RI) spatiales dans le cadre de bibliothèques numériques liées au patrimoine culturel local. L'approche proposée (implémentée dans le prototype PIV[1]) est construite autour d'une analyse sémantique de tels corpus et de requêtes écrites en texte libre. Nous présentons la méthodologie d'annotation sémantique pour l'indexation automatique et le géo-référencement de documents textuels. Un cas d'étude permet ensuite d'évaluer le processus de RI spatiale proposé. Dans ce cas d'étude, nous comparons notre approche avec les approches classiques basées sur les statistiques en utilisant d'abord des requêtes purement spatiales, puis des requêtes plus générales. Le principal résultat de ces premières expérimentations montre que la combinaison de l'approche spatiale et de l'approche statistique basée mots clés, améliore de manière significative la précision, notamment dans le cas de requêtes « générales ».*

ABSTRACT. *This paper deals with Information Extraction and Retrieval in a Geographic oriented Digital Libraries environment. The proposed approach (implemented within PIV[1] prototype) is based on a semantic analysis of digital corpora and free text queries. First, we present requirements and a methodology of semantic annotation for automatic indexing and geo-referencing of text documents. Then we report on a case study where the spatial-based IR process is evaluated and compared to classical (statistical-based) IR approaches using first, pure spatial queries and then, more general ("realistic") ones containing both spatial and thematic scopes. The main result in these first experiments shows that combining our spatial approach with a classical (statistical-based) IR one improves in a significant way retrieval accuracy, namely in the case of "realistic" queries.*

MOTS-CLÉS : *Extraction d'Information, Recherche d'Information, Dimension Géographique de l'Information, Bibliothèques Numériques.*

KEYWORDS: *IE, Information Retrieval, Geographic Information Scope, Digital Libraries.*

[1] PIV: Le projet Pyrénées Itinéraires Virtuels est soutenu par la Communauté d'Agglomération Pau Pyrénées (CDAPP) et la Médiathèque Intercommunale à Dimension Régionale (MIDR).

1. Introduction

Une étude faite sur le moteur de recherche Excite montre qu'une requête sur cinq effectuée sur celui-ci est liée à la géographie (Sanderson et al., 2004). Notre contribution se situe dans l'amélioration des systèmes de gestion documentaire (GED, etc.) existants et propose d'ajouter aux services basiques de ces outils des services dédiés au traitement de l'information géographique (projet PIV). L'information géographique contenue dans les corpus de ces bibliothèques est composée d'entités spatiales (Lesbegueries et al., 2006), temporelles et thématiques. "Les instruments de musique dans les environs de Laruns au XIXème siècle" est un exemple d'entité géographique complète : "Instruments de musique" est l'entité thématique, "dans les environs de Laruns" est l'entité spatiale et "au XIXème siècle" l'entité temporelle.

Prenons l'hypothèse que pour enclencher un processus de recherche géographique, l'entité spatiale doit être explicite tandis que l'entité temporelle peut être implicite ou être exprimée ailleurs dans le texte (dans ce cas elle peut être associée à plusieurs entités spatiales) et l'entité thématique (ou phénomène) peut être inexistante. Par conséquent, l'analyse de l'information spatiale est nécessaire pour faire un traitement géographique approfondi.

Notre modèle spatial supporte des Entités Spatiales Absolues ou Relatives (ESA ou ESR). Les entités spatiales nommées comme "la ville de Bayonne" sont des entités nommées connues. Nous les définissons dans notre modèle comme des ESA. Les entités spatiales complexes comme "à 10 km au sud de la frontière franco-espagnole" doivent être interprétées et nécessitent donc un traitement particulier intégrant du raisonnement spatial. Ces entités sont définies comme des ESR. Nous associons à chaque ESR au moins une relation (d'adjacence, d'inclusion, de distance ou d'orientation) pour une définition récursive (Lesbegueries et al., 2006 et 2006b).

La différence de notre approche par rapport à d'autres comme celle du projet SPIRIT (Jones et al., 2004) ou de GIPSY (Woodruff et al., 1994) vient du raisonnement spatial réalisé pour l'interprétation et l'indexation des ESAs et des ESRs. Par exemple, le système de SPIRIT ne marque que des entités équivalentes à nos ESAs (plutôt des adresses) dans des documents web. Une autre spécificité vient de la granularité des unités d'information gérées : des paragraphes de texte provenant de corpus numérisés spécifiques à un domaine (patrimoine culturel pyrénéen) dans notre cas et des pages web dans le cas de SPIRIT. Dans l'approche proposée, une interprétation fine de l'information spatiale et le processus de marquage sont appliqués à la fois à l'étape d'indexation des unités de texte et à l'étape de l'interprétation de la requête. Comme nous travaillons sur des collections spécifiques, limitées en taille et relativement stables (contrairement aux pages web par exemple), notre approche (plus fouillée) semble appropriée. Ainsi, le coût d'une telle indexation spatiale demeure raisonnable. Les requêtes en texte libre sont interprétées dynamiquement selon le même procédé. Les index détaillés des entités spatiales permettent une recherche d'information plus précise.

La section 2 présente des propositions relatives au traitement de l'information spatiale. Ensuite, la section 3 décrit l'approche retenue dans PIV. Enfin, la section 4 évalue et compare l'approche spatiale PIV à une approche de RI classique, puis propose de combiner ces deux approches.

2. L'information spatiale

Plusieurs études ont décrit la manière particulière d'exprimer l'information spatiale dans le langage écrit. Un lieu correspond à une catégorie (parcelles de terrain, étendues d'eau et lieux habités) et est associé à une frontière naturelle ou artificielle (Borillo, 1998). Dans (Vandeloise, 1986), Vandeloise parle du concept de couple site/cible. En effet, nous pouvons comprendre la phrase 1 suivante parfaitement alors que la phrase 2 paraît non appropriée (Lesbegueries et al., 2006b):

– phrase 1 : *la voiture est près de la maison*

– phrase 2 : *la maison est près de la voiture*

Donc, dans le langage écrit, une information spatiale peut être définie en référençant un lieu bien connu.

La communauté SIG (Système d'Information Géographique) modélise l'information comme un ensemble de données géo-référencées : une entité spatiale (ES) et toutes les relations spatiales ont été définies à partir de trois primitives : la direction, la distance et des opérations booléennes ensemblistes. (Freeman, 1975) a défini différentes relations spatiales (à gauche de, à droite de, au-dessus de, en-dessous de, loin, près, à côté de, dans, etc.). Des travaux récents se sont intéressés au raisonnement spatial qualitatif. Egenhöfer et Franzosa (Egenhofer et al., 1991) ont développé un modèle souvent cité depuis, composé de relations topologiques entre des objets spatiaux. (Clementini et al., 1994) ont étendu ce modèle en prenant en compte la dimension des intersections. Ces travaux ont défini des relations spatiales différentes et les opérations associées. Enfin, les travaux autour des bibliothèques numériques (Hill, 2000) présentent les gazetteers comme une forme particulière de dictionnaire de lieux nommés.

Comme les approches d'extraction et de recherche d'information sont assez génériques, la gestion précise de l'information spatiale est encore un grand défi. L'approche sémantique semble être un moyen pertinent de gérer l'information spatiale dans les systèmes d'EI et de RI.

Dans l'information textuelle, la chaîne de traitement sémantique utilisée pour extraire des marqueurs spatiaux est composée de quatre grandes étapes (Abolhassani, 2003) : (1) la lemmatisation pour segmenter les mots ; (2) l'analyse lexicale et morphologique pour la reconnaissance des mots ; (3) l'analyse syntaxique, basée sur des grammaires, afin de trouver les relations entre les mots ; (4) enfin l'analyse sémantique pour réaliser une interprétation plus spécifique sur les syntagmes retenues (Charnois et al., 2003; Wildocher et al., 2005).

3. Le projet PIV

3.1 Vue globale du système

L'objectif principal vise la valorisation de corpus fortement territorialisés auprès d'utilisateurs non-experts (touriste, scientifique ou scolaire). La Figure 1 représente une vue globale du système PIV comportant les deux processus principaux d'extraction/indexation et de recherche d'information spatiale.

La partie extraction (cf. Fig.1) est découpée en quatre étapes. La première (1) concerne la collecte d'ouvrages numérisés relatant du patrimoine culturel pyrénéen. La seconde (2) supporte une analyse linguistique puis sémantique afin d'extraire des ESs sous forme d'instances du modèle spatial de PIV. La troisième (3) s'appuie sur des ressources géographiques (communes, lieux-dits, routes, pics, vallées, …) afin de valider les ESs détectées à l'étape précédente. La dernière étape (4) propose le calcul de représentations spatiales géo-référencées. Ainsi, les résultats obtenus sont soit des ESs absolues (i.e. "le village de Laruns") soit des ESs relatives (i.e. "les environs de Laruns").

La partie recherche d'information est basée sur une analyse similaire de la requête (6) puis, d'un appariement spatial. Cet appariement calcule les surfaces d'intersection (7) entre les représentations spatiales correspondant à la requête et celles contenues dans les index. Il est alors nécessaire d'extraire les fragments de documents jugés pertinents (8) et, enfin, de les présenter à l'utilisateur (9).

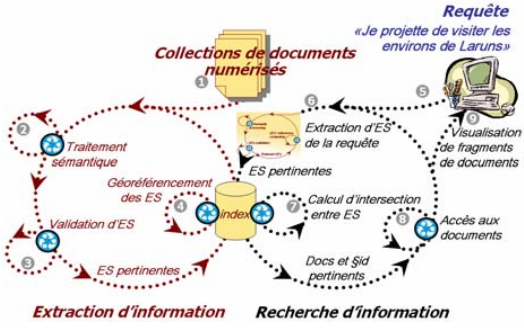

Figure 1. *Schéma synoptique des processus d'EI et de RI*

3.2 Modèle spatiale unifié

Ce modèle, présenté dans (Lesbegueries et al., 2006 et 2006b), applique l'hypothèse linguistique citée plus haut : une ES est définie récursivement à partir d'autres ESs et de relations spatiales. Le principe site/cible (Vandeloise, 1986) peut en effet être décrit d'une manière récursive. Par exemple, l'ES "au nord de la ligne Biarritz-Pau" est d'abord définie par les repères ou sites "Biarritz" et "Pau" qui sont

des lieux supposés connus. Le terme "ligne" crée un objet géométrique liant les premiers repères et coupant l'espace en deux. Finalement une relation d'orientation détermine l'un des deux espaces créés par la ligne. Une ES peut donc être une ES absolue (ESA "Biarritz") ou une ES relative (ESR définie par une relation spatiale et une autre ES). Les relations spatiales définies pour une ESR sont l'adjacence ("près de Laruns"), l'orientation ("à l'ouest de Laruns"), la distance ("à 10 km de Laruns"), l'inclusion ("au centre de Laruns") et la forme géométrique ("le triangle Laruns Arudy Mauléon").

3.3 EI spatiale

La chaîne de traitement linguistique et sémantique de PIV présentée dans (Lesbegueries et al., 2006 et 2006b) supporte l'extraction et l'indexation d'information spatiale. Cette chaîne traite des sources d'information hétérogènes (des journaux, des livres, des notices descriptives de cartes postales et de lithographies) et alimente des index. Nous l'utilisons également pour séparer les entités spatiales des entités thématiques dans la requête pour la RI (voir section Etude de cas).

Nous adoptons une approche de parcours active du corpus. Contrairement aux approches standards de traitement automatique de la langue, notre chaîne de traitement linguistique est appliquée localement aux entités spatiales. Les ESAs (noms de villages, de forêts, etc.) sont d'abord détectées et marquées en se basant sur un lexique composé notamment d'introducteurs spatiaux. Puis, les ESRs sont construites à partir de ces ESAs.

Un *étiqueteur* et un *sectionneur* (splitter) parcourent le flux de texte et insèrent des marqueurs de structures logiques et des séparateurs de mots avec leur lemme. Les entités spatiales candidates sont détectées dans une seconde étape : tous les termes commençant par une majuscule et précédés d'un introducteur d'entité spatiale appartenant au lexique prédéfini ("dans", "proche de", etc.) sont marqués. Puis, un *étiqueteur syntaxique* (POS tagger) classe ces entités nommées (noms propres, etc.). Une analyse basée sur les grammaires DCG (Definite Clause Grammar) permet ensuite l'interprétation des syntagmes extraits (inclusion, adjacence, distance par rapport à une autre entité spatiale, etc.). L'expression "près de Laruns" est interprétée comme une entité spatiale relative ESR, elle-même définie par la relation d'adjacence relative à une entité spatiale absolue ESA ("Laruns"). Enfin, une étape de validation fait appel à des services externes (gazetteers) pour confirmer chaque ESA candidate. Toutes les ESR candidates associées à une ESA non valide sont automatiquement supprimées. Une représentation géolocalisée de chaque ES validée est ensuite ajoutée aux index.

3.4 RI spatiale basée sur l'intersection des ESs

La recherche dans le corpus ainsi indexé commence par une extraction des

informations spatiales exprimées dans la requête (traitement similaire de détection d'ESs). Elle est suivie d'un appariement entre les ESs de la requête et celles contenues dans les index. Cet appariement est basé sur le calcul d'intersections entre les zones géo-référencées (boites englobantes) correspondant aux ESs de la requête et des index (Sallaberry et al., 2006).

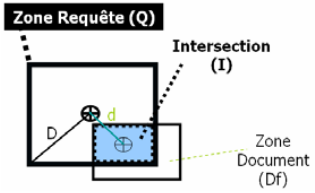

Figure 2. *Calcul de pertinence d'un document sélectionné*

Pour chaque requête, on est capable de calculer la pertinence d'un document de la collection en calculant d'abord le *Df precision* qui est égal au rapport de la surface d'intersection entre la requête (*I surface*) et le document (*Df surface*) (Figure 2) :

$$Df\ precision = \frac{I\ surface}{Df\ surface}$$

Puis le rapport avec la requête (*Df significance)* : $Df\ significance = \frac{I\ surface}{Q\ surface}$

Et le rapport de distance : $Df\ distance = \frac{d}{D}$

Ainsi, la valeur de pertinence *Df score* est calculée comme suit :

$$Df\ score = \frac{\left(Df\ precision + Df\ significance\right)}{\left(2 + Df\ distance\right)} \quad (1)$$

Plus les centroïdes de I et Q sont proches, plus la pertinence du document est grande. Le SGBD[2] XML et que le GIS[3] utilisés supportent ces opérations.

4. Etude de cas

Dans cette section, nous évaluons l'approche de RI spatiale de PIV basée sur les index spatiaux construits précédemment. Les résultats de PIV sont comparés à ceux obtenus par l'approche classique de RI basée mots clés sur la même collection et le

[2] *eXist*. http://exist.sourceforge.net
[3] *PostGIS*. http://postgis.refractions.net

même ensemble de requêtes test. L'approche classique utilisée est décrite dans la section suivante.

4.1 L'approche de RI classique

L'approche de recherche d'information classique est basée sur la notion de « sac de mot » (Baeza-Yates, 1999). Dans ces approches, les documents sont indexés en utilisant une indexation classique des termes. Elle consiste d'abord à sélectionner les mots simples occurrant dans les documents, puis à les lemmatiser (Porter, 1980) et enfin à enlever les mots vides (stoplist). Nous avons utilisé une liste de mots vides et un lemmatiseur de la langue Française de la famille Snowball[4]. Un poids *Wtd(t,d)* est ensuite assigné à chaque terme t d'un document d_j suivant la formule (2) :

$$Wtd(t_j, d_j) = \frac{\frac{2.tf_{ij}.\log(N - n_i + 0.5)}{(n_i + 0.5)}}{2.(0.25 + 0.75\ .\ dl_j / avg\_dl) + tf_{ij})} \qquad (2)$$

Où $tf_{i,j}$ représente la fréquence du terme t_i dans le document d_j, n_i est le nombre de documents contenant le terme t_i et N le nombre total de documents dans la collection. dl_j représente la taille du document d_j et *avg_dl,* la taille moyenne de document dans la collection. Cette méthode de pondération, qui est une amélioration de la formule TF.IDF est introduite pour atténuer l'impact négatif des documents longs lors de la phase de recherche (Robertson, 1999). Ceci est bien adapté aux paragraphes de tailles variées de notre collection. Le même processus d'indexation est appliqué aux requêtes.
Un modèle de recherche vectoriel (Boughanem, 2001) est ensuite utilisé pour la phase de recherche : pour une requête q donnée, le produit scalaire (Inner product) entre le vecteur de la requête et ceux correspondants à chaque document d_j dans la collection est appliqué pour calculer les scores de pertinence *Rel (q, d_j)* :

$$Rel(q, d_j) = \sum_{k=1}^{|q|} Wtq(t_k, q).Wtd(t_{k,} d) \qquad (3)$$

Ce score de pertinence est utilisé pour déterminer le classement du document d_j (ranking) dans la liste finale des documents sélectionnés en réponse à la requête q.

4.2 Collection de test

Le corpus utilisé pour évaluer le système PIV est fourni par la Médiathèque MIDR. Il contient 10 livres OCRisés traitant du patrimoine culturel Pyrénéen et datant des XIXème et XXème siècles. Les livres sont découpés en paragraphes constituant 10000 unités documentaires. Des bibliothécaires ont créé 12 requêtes

[4] http://snowball.tartarus.org/texts/introduction.html.Ricardo

dont 8 traitent de la dimension spatiale seulement et les quatre autres traitent de la dimension spatiale et thématique. Une requête spatiale peut faire référence à des entités spatiales absolues (ESA) ou relatives (ESR). Une requête spatiale et thématique telle « instruments de musique dans les environs de Laruns au XIXème siècle » supporte aussi bien les entités ESA/ESR ("environs de Laruns") que d'autres entités non spatiales ("instruments de musique", "XIXème siècle"). Le prototype PIV a trouvé 9835 ESs candidates dans les dix livres (le traitement d'un livre de 200 pages prend cinq minutes.). Nous avons aussi annoté manuellement les entités spatiales : comme pour la campagne CLEF6, les participants marquent manuellement les ESs trouvées dans l'ensemble du corpus pour avoir des résultats de référence.

4.3 Evaluation de l'approche de RI spatiale

Nous avons soumis les huit requêtes spatiales au système PIV et nous avons comparés les premiers documents restitués (top 5, 10 et 15) aux jugements manuels. Les résultats sont donnés en Table 1 où Avg représente la précision moyenne calculée sur toutes les requêtes utilisées et P@5, P@10 et P@15 désignent respectivement la précision aux points top 5, top 10 et top 15. La dernière colonne, Nombre de réponses, représente le nombre total de documents trouvés (moyenne sur toutes les requêtes).

Toutes les requêtes	P@5	P@10	P@15	Nombre de réponses
A) Approche Spatiale (PIV)				
Avg	0.78	0.81	0.73	637
B) Approche classique				
Avg	0.50	0.43	0.40	252

Table 1. *Résultats de PIV et de l'approche classique sur les requêtes spatiales*

On peut remarquer que l'approche spatiale PIV (Table 1-A) donne une précision de 78% dans les cinq premiers documents (top 5) et 81% dans les dix premiers. Quand les mêmes requêtes sont soumises au système classique, les résultats se dégradent de manière significative (Table 1-B). La raison est que pour une requête spatiale telle "près de Laruns", l'approche classique (basée sur une comparaison mot-mot) ne trouve jamais les documents traitant d'autres villes comme "Eaux-Bonnes" ou "Louvie-Soubiron", qui sont localisées dans le voisinage de "Laruns". Notre approche, en extrayant les ESs à partir du texte des documents et des requêtes, propose une comparaison intégrant une interprétation de la sémantique de ces ESs pour récupérer les documents pertinents. Ce résultat est aussi confirmé si nous considérons le nombre de documents restitués : en moyenne, 637 documents sont trouvés par l'approche spatiale pour toutes les requêtes ; alors que seulement 252 sont sélectionnés par l'approche classique.

4.4 Evaluation de la RI thématique et spatiale

Nous regardons l'utilisation de requêtes plus générales, contenant une dimension spatiale et thématique. La Table 2 donne les résultats obtenus par l'approche spatiale (A) et l'approche classique (B). On peut remarquer que les résultats sont sensiblement dégradés pour l'approche PIV : seulement 15% de précision aux cinq premiers documents restitués pour l'approche PIV alors que l'approche classique arrive à une précision de 48%. Une analyse plus fine de ces résultats montre que des documents pertinents sont trouvés mais ne sont pas classés au début. En fait, l'approche PIV n'est pas adaptée pour le classement dans le cas de requêtes générales (espace + thème).

Toutes les requêtes	P@ 5	P@10	P@15	Nombre de réponses
A) Approche spatiale (PIV)				
Avg	0.15	0.18	0.18	1154
B) Approche classique				
Avg	0.48	0.39	0.36	331

Table 2. *PIV et l'approche classique sur les requêtes spatiales + thématiques*

Comme dans le premier cas, les mêmes requêtes sont soumises au système classique. Les résultats (Table 2-B) sont clairement plus favorables à l'approche classique. Par exemple, pour l'approche classique, le système arrive à une précision moyenne de 48% aux 5 premiers documents et 36% aux 15. L'équivalent pour l'approche PIV est de seulement 15% et 18% respectivement. On peut aussi noter la différence dans le nombre des documents sélectionnés : 1154 documents pour PIV et 331 pour l'approche classique. Ceci est du au fait que PIV cherche toutes les ESs absolues et relatives jugées pertinentes en réponse à la dimension spatiale de la requête sans prendre en compte la pertinence globale par rapport à la requête entière.

4.5 Combinaison des approches spatiales et classiques

Les résultats précédents suggèrent que d'un côté, l'approche spatiale est bien adaptée à la recherche de documents traitant de cette dimension mais l'est moins quand il s'agit de classer les documents pertinents dans les requêtes générales (espace + thème). L'approche classique, quant à elle, manque d'exhaustivité quand elle traite des requêtes spatiales, mais surclasse l'approche PIV dans le cas de requêtes générales. Nous avons pensé à combiner les deux approches pour améliorer le taux de précision des requêtes traitant de l'espace et du thème. De plus, le fait qu'une unité documentaire correspond à un paragraphe augmente la probabilité que les informations spatiales et thématiques soient sémantiquement liées. L'idée est de subdiviser une requête en deux sous-requêtes (Figure 3) ; la *sous-requête spatiale* et

la *sous-requête thématique*. La *sous-requête spatiale* contient les ESs identifiées par la chaîne de traitement linguistique. La *sous-requête thématique* contient les termes restants de la requête (temps, événement). Comme schématisé dans la Figure 3, "les environs de Laruns" et "instrument de musique du XIXème siècle" représentent respectivement la *sous-requête spatiale* et la *sous-requête thématique* de la requête exemple.

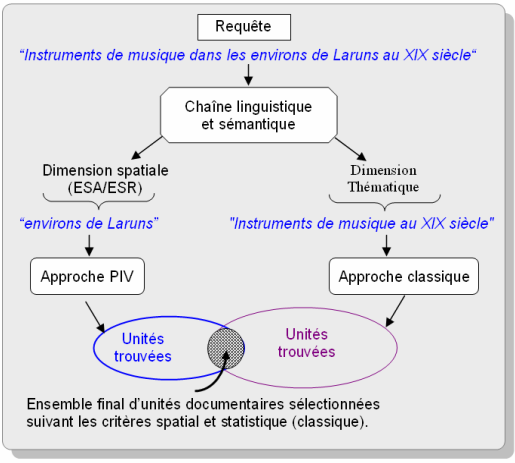

Figure 3. *Combinaison des approches spatiale et thématique*

Les deux sous-requêtes sont ensuite soumises au système supportant l'approche appropriée. Le résultat final est ensuite construit en faisant une intersection des deux ensembles de documents sélectionnés par PIV et l'approche classique. Le classement final est basé sur celui obtenu par PIV : chaque document classé dans l'ensemble de PIV est ajouté à l'ensemble final s'il est également classé dans l'ensemble retourné par le système classique. Le détail des résultats obtenus en utilisant les requêtes précédentes (espace + thème) suivant cette stratégie sont donnés dans la Table 3.

Toutes les requêtes	P@5	P@10	P@15	Nombre de réponses
	A) Combiner les résultats de l'approche spatiale et classique			
Avg	**0.70**	**0.50**	**0.43**	**25.75**

Table 3. *Résultats de la combinaison de PIV et de l'approche classique*

Les résultats confirment l'hypothèse de départ : combiner l'approche spatiale basée sur l'extraction et la comparaison des entités spatiales absolues et relatives

avec l'approche thématique basée sur les statistiques permet d'améliorer les performances du système en classant plus de documents pertinents. Par exemple, aux cinq premiers documents, la précision atteint 70% quand les approches classique est spatiale sont combinées ; alors qu'elle n'est respectivement que de 48% et 15% quand les deux approches sont utilisées séparément. Cependant, on peut noter le nombre réduits de documents sélectionnés à cause de la méthode "prudente" de combinaison adoptée (intersection simple) : par exemple, dans le cas d'une de nos requêtes, l'approche combinée sélectionne seulement 4 documents alors que l'approche classique en retourne 233 et l'approche PIV 724. L'amélioration de la précision au niveau des premiers documents restitués s'accompagne ainsi d'une diminution au niveau du rappel.

Un domaine ouvert concerne donc le problème de fusion des deux ensembles de résultats pour optimiser aussi bien la précision dans les premiers documents que le rappel. Ceci nécessite l'étude d'opérateurs plus complexes (union, avec intégration pondérée de la pertinence spatiale et statistique, par exemple) que la simple intersection utilisée ici.

5. Conclusion

Dans ce papier, nous avons proposé une approche d'extraction et de recherche d'information spatiale dans les bibliothèques numériques. Un travail sur le raisonnement spatial a permis de concevoir une méthode de structuration de telles informations contenue dans les documents. Cette méthode a été validée par l'implémentation d'un prototype de RI spatiale.
L'évaluation de l'approche proposée a permis de mettre au clair : 1) la supériorité de notre approche par rapport à l'approche classique dans le cas de requête spatiales pures ; 2) la nécessité de combiner notre approche avec l'approche statistique classique dans le cas de requêtes plus générales (traitant aussi bien de l'espace que d'autres thèmes).

Globalement, notre contribution permet de compléter les approches traditionnelles utilisées dans les bibliothèques numériques. Sur le plan opérationnel, le système PIV utilise des services web et supporte le format XML (schémas, représentation d'ESs, index, etc.). Les services web de PIV peuvent donc être facilement intégrés dans les systèmes de gestion électronique de documents ou de bibliothèques existants.

Une perspective envisageable à ce travail concerne la combinaison des résultats des approches de PIV et classique. En effet, dans l'étude de cas, nous nous sommes contentés d'utiliser un opérateur relativement simple, l'intersection, pour fusionner les deux ensembles de résultats. Il serait intéressant d'étudier, l'utilisation d'autres opérateurs plus complexes et leur impact aussi bien sur la précision que sur le rappel.

6. References

Abolhassani, M., Fuhr, N., Govert; N. (2003). Information Extraction and Automatic Markup for XML documents, *Intelligent Search on XML Data, LNCS Springer*, p. 159–178.

Baeza-Yates, R. A., Ribeiro-Neto., B. A. (1999). Modern Information Retrieval. *ACM Press / Addison-Wesley.*

Borillo, A. (1998). L'espace et son expression en français. *L'essentiel. Ophrys.*

Boughanem, M., Chrisment, C., Tmar, M. (2001). Mercure and MercureFiltre Applied for Web and Filtering Tasks at TREC-10. In *Proceeding of TREC*.

Charnois, T., Mathet, Y., Enjalbert, P., Bilhaut, F. (2003). Geographic reference analysis for geographic document querying. *Workshop on the Analysis of Geographic References, Human Language Technology Conference (NAACL-HLT).*

Clementini, E., Sharma, J., and Egenhofer, M. (1994). Modeling topological spatial relations: Strategies for query processing. *Computers and Graphics,* p. 815-822.

Egenhofer, M. J., Franzosa, R.D. (1991). Point-Set Topological Relations. *International Journal for Geographic Information Sytems,* 5(2):161-174.

Freeman, J. (1975). The Modelling of Spatial Relations. *Computer Graphics and Image Processing,* 4:156-171.

Hill, L. (2000). Core elements of digital gazetteers: Place names, categories, and footprints. In *ECDL '00: Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries,* p. 280–290. Springer-Verlag.

Jones, C.-B., Abdelmoty, A.-I., Finch, D., Fu, G., Vaid, S. (2004). The Spirit Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. *Third International Conference - Geographic Information Science, Adelphi, Usa,* pp. 125 – 139.

Lesbegueries, J., Gaio, M., Loustau, P., and Sallaberry, C. (2006). Geographical information access for non-structured data. *ACM SAC - Advances in Spatial and Image based Information Systems track.*

Lesbegueries, J., and Loustau, P. (2006b). Extraction et interprétation d'information géographique dans des données non-structurées. *RIAO*

Porter M. 1980, An algorithm for suffix stripping, Program, 14(3) pp 130−137. http://snowball.tartarus.org/texts/introduction.html.Ricardo

Robertson, S.E., Walker, S. , Hancock-Beaulieu, M., Gatford, M., Payne A. (1995). Okapi at TREC-4, 1995. In *Proceeding of TREC.*

Sallaberry, C., Etcheverry, P., and Marquesuzaà, C. (2006). Information Retrieval and Visualization Based on Documents' Geospatial Semantics. *4th IEEE International Conference on Information Technology: Research and Education, ITRE.*

Sanderson, M. and Kohler, J. (2004). Analyzing geographic queries. In *Proceedings of the Workshop on Geographic Information Retrieval, SIGIR,* www.geo.unizh.ch/ ~rsp/gir/

Vandeloise, C. (1986). L'espace en français. *Travaux Linguistiques. Seuil.*

Wildocher, A., Bilhaut, F. (2005). La plate-forme linguastream : un outil d'exploration linguistique sur corpus. In *Actes de la 12e Conférence Traitement Automatique du Langage Naturel.*

Woodruff, A.G., Plaunt, C. (1994). GIPSY: Automated Geographic Indexing of Text Documents. *Journal of the American Society for Information Science,* 45:9:645-655.