
Cadre d'évaluation de systèmes de recherche d'information géographique

Apport de la combinaison des dimensions spatiale, temporelle et thématique

Damien Palacio* — **Guillaume Cabanac****
Christian Sallaberry* — **Gilles Hubert****

* Université de Pau et des Pays de l'Adour, LIUPPA ÉA 3000
Avenue de l'Université, BP 1155, F-64013 Pau cedex

** Université de Toulouse, IRIT UMR 5505 CNRS
118 route de Narbonne, F-31062 Toulouse cedex 9

prénom.nom@univ-pau.fr, prénom.nom@irit.fr

Catégorie chercheur

RÉSUMÉ. Les moteurs de recherche communément utilisés exploitent les termes contenus dans les documents pour répondre aux besoins d'information des individus. De telles approches s'avèrent limitées dans des contextes spécifiques tels que la gestion de collections spécialisées (notamment patrimoniales) ou la mise en œuvre de critères d'interrogation ciblés (notamment multidimensionnels). Dans cet article, nous considérons le cas des systèmes de recherche d'information (SRI) géographiques exploitant les dimensions spatiale, temporelle et thématique. Notre contribution est double, en proposant un cadre d'évaluation de tels systèmes que nous exploitons pour tester l'hypothèse suivante : la combinaison des dimensions spatiale et temporelle avec la dimension thématique améliore la qualité des résultats.

ABSTRACT. Common search engines process users' queries, i.e., information needs, by extracting terms from documents. Such approaches are limited regarding particular contexts, such as specialized collections (e.g., cultural heritage collections) or specific retrieval criteria (e.g., multidimensional criteria). In this paper, we consider Geographic Information Retrieval Systems (GIRS) exploiting the spatial, temporal, and topical dimensions. Our contribution is twofold as we propose a GIRS evaluation framework for testing the following assumption: combining spatial and temporal dimensions along with the topical dimension improves GIRS effectiveness.

MOTS-CLÉS : Recherche d'information géographique, évaluation, combinaison de résultats.

KEYWORDS: Geographic Information Retrieval, evaluation, retrieval combination.

1. Introduction

La numérisation de documents progresse, produisant un volume croissant de documents électroniques. Tandis que certains projets visent simplement la création de versions numérisées de documents, des efforts centrés sur des domaines spécifiques ont des objectifs bien plus ambitieux. Par exemple, afin d'exploiter leurs contenus, les documents sont annotés et indexés conformément à des modèles de données dédiés à la description de domaines particuliers. Ainsi, les fonds documentaires patrimoniaux mobilisent de gros efforts de numérisation. Leur valorisation se fait ensuite via des outils de gestion documentaire classiques (notices descriptives, catalogues thématiques) intégrant généralement des moteurs de recherche plein-texte.

Dans ce contexte, le projet PIV¹ « Pyrénées Itinéraires Virtuels » (Gaio *et al.*, 2008) consiste à gérer un fonds documentaire de versions électroniques de documents du XIX^e siècle consacré aux Pyrénées et constitué de journaux, romans et récits de voyages. Il s'agit d'un fonds stable (peu de suppressions et de modifications, des insertions régulières de nouveaux documents) et de petite taille (quelques dizaines de milliers de documents). Ce type de ressource est encore très peu connu et son usage limité aux centres des archives et bibliothèques locales. C'est la raison pour laquelle les collectivités locales souhaitent valoriser ces ressources en les diffusant largement et en offrant des services de recherche adaptés.

Des travaux ont quantifié la proportion des requêtes géographiques au sein des sessions de recherches issues de moteurs généralistes. Ainsi, pour Excite (Sanderson *et al.*, 2004), AOL (Gan *et al.*, 2008) et Yahoo (Jones *et al.*, 2008), cette proportion varie entre 12,7 % et 18,6 %. Or, bien que les moteurs généralistes restituent déjà de bons résultats pour des recherches par mots-clés, Kanhabua *et al.* (2008) ont observé que, sur des corpus aussi spécifiques, la précision des recherches géographiques est faible. De fait, l'utilisateur passe beaucoup de temps à explorer les documents restitués afin de ne retenir que ceux qui satisfont réellement son besoin. Par exemple, la requête « *les années 1810* » avec un moteur de recherche classique restitue des documents contenant « *1810* » et non « *1811* », « *1812* », etc. De même, la requête « *Anglet* » avec un moteur de recherche classique restitue des résultats contenant « *Anglet* » et non « *le golf de Chiberta* », « *la plage des Cavaliers* », etc. Un moyen d'améliorer l'efficacité des moteurs de recherche est alors d'y inclure la prise en compte des aspects géographiques. Nous reprenons la théorie selon laquelle l'information géographique comporte trois dimensions : spatiale, temporelle et thématique (Gaio, 2001). L'extrait de texte « *Les villes et les châteaux fortifiés dans le bassin aquitain au XIII^e* » illustre bien ce triptyque.

Pour pallier les insuffisances des moteurs de recherche généralistes, dans le cadre du projet PIV, nous avons développé trois chaînes de traitement dédiées à l'indexation et à la recherche d'information spatiale, temporelle et thématique. Un prototype, correspondant à chaque chaîne de traitement, extrait et indexe des informations is-

1. Le projet PIV est soutenu par la Communauté d'Agglomération Pau Pyrénées.

sues de documents textuels et propose un moteur de recherche qui, sur des critères spatiaux (Gaio *et al.*, 2008), temporels (Le Parc-Lacayrelle *et al.*, 2007) ou thématiques (Sallaberry *et al.*, 2007), restitue des paragraphes de documents (récits de voyage, par exemple) associés à un score de pertinence. Ces travaux se situent à la croisée de plusieurs disciplines : Traitement Automatique des Langues Naturels (TALN), Systèmes d'Information Géographiques (SIG), Recherche d'Information (RI) et Recherche d'Information Géographique (RIG).

La contribution de cet article vise la conception et la mise en place d'un cadre d'évaluation de SRI géographique. Le cadre d'évaluation que nous proposons a) est destiné à évaluer les SRI spatiaux, temporels, thématiques et toute combinaison de tels SRI ; b) il met en place une collection de test en langue française adaptée à la RIG et, à moyen terme, c) il pourrait mener à la proposition d'une tâche spécifique dans le cadre d'une campagne d'évaluation de type GeocLEF (Gey *et al.*, 2006).

L'article est structuré comme suit. La section 2 présente une synthèse de l'état de l'art relatif à l'évaluation de SRI, y compris géographiques. La section 3 est consacrée à la description du cadre d'évaluation de SRI géographique que nous proposons. La section 4 détaille une étude de cas : nous évaluons le SRI spatial, le SRI temporel et le SRI thématique du système PIV ainsi que différentes combinaisons de ces SRI. Notre hypothèse est que la combinaison des trois dimensions donne des résultats plus pertinents que chacune prise indépendamment. Cette hypothèse est vérifiée au travers de l'analyse des résultats d'évaluation des SRI. La section 5 fait état des SRI géographiques de la littérature que l'on pourrait également évaluer avec le cadre proposé dans cet article. Enfin, la section 6 conclut l'article et discute nos perspectives de recherche.

2. Adéquation des cadres d'évaluation existants pour la RI géographique

Le domaine de la RI est caractérisé par une longue tradition d'évaluation, notamment au travers du cadre TREC (Voorhees *et al.*, 2005) qui permet l'évaluation des SRI au regard de la dimension thématique. Par ailleurs, la dimension temporelle a également fait l'objet du cadre d'évaluation TEMPEVAL (Verhagen *et al.*, 2009). Bucher *et al.* (2005) ont proposé de considérer deux dimensions simultanément : spatiale et thématique. Cette proposition se retrouve dans la tâche GeocLEF (Gey *et al.*, 2006) du cadre CLEF (Peters, 2001). Ce dernier a notamment permis, à partir de requêtes géographiques, l'évaluation de SRI thématiques classiques en RI tels que Lemur (Ogilvie *et al.*, 2001), Lucene (Gospodnetić *et al.*, 2005) et Terrier (Ounis *et al.*, 2005), comme rapporté dans (Perea-Ortega *et al.*, 2008).

Les travaux existants (présentés en section 5) ont tout au plus été évalués du point de vue de la taille des index générés et du temps de réponse des SRI géographiques. Ces évaluations quantitatives gagneraient à être mises en perspective avec des évaluations qualitatives. Or, à notre connaissance, il n'existe pas de cadre d'évaluation des trois dimensions de l'information géographique d'un point de vue qualitatif. Il est

donc impossible de comparer les moteurs de recherche qui s'efforcent de les traiter simultanément. Pour répondre à ce besoin nous proposons donc, dans la section suivante, un cadre expérimental permettant d'évaluer la RI géographique.

3. Proposition d'un cadre d'évaluation dédié à la RI géographique

Le cadre expérimental proposé s'attache à capitaliser le savoir-faire existant (issu notamment de TREC et GeoCLEF) tout en intégrant les spécificités manquantes relatives à l'information géographique. Aussi, la section 3.1 détaille la constitution d'une collection de test couvrant les trois dimensions, puis la section 3.2 expose l'analyse des résultats de SRI permettant de comparer leur efficacité.

3.1. Constitution d'une collection de test pour évaluer la RI géographique

Dans la littérature, notamment à TREC (Harman, 2005), une collection de test comprend trois volets :

1) un ensemble de n « *topics* » formulés par des individus, où *topic* est le terme TREC désignant un besoin d'information. Chaque *topic* est au moins caractérisé par un titre, une description et une narration du besoin. Buckley *et al.* (2000) montrent qu'au moins 25 *topics* sont nécessaires pour réaliser des analyses statistiques pertinentes. Notons cependant que le standard de TREC est à 50 *topics*.

2) le *corpus* regroupant plusieurs documents, certains étant pertinents pour les *topics* proposés. Les corpus TREC comprennent plusieurs centaines de milliers de documents au moins (Voorhees *et al.*, 2005).

3) les « *qrels* », terme TREC désignant les *jugements de pertinence*, associant à chaque *topic* l'ensemble des documents pertinents. Étant donné que le corpus est trop volumineux pour être exhaustivement analysé dans le but d'identifier les *qrels*, TREC recourt à la technique du *pooling*. Ainsi, pour chaque *topic*, un *pool* de documents est constitué à partir des 100 premiers documents restitués par chacun des systèmes participant à la campagne d'évaluation, les doublons sont supprimés (opération d'union ensembliste). L'hypothèse est que le nombre et la diversité des SRI contribuant au *pool* permettront de trouver un maximum de documents pertinents. Enfin, un individu appelé « *assesseur* » examine chaque document du *pool* afin d'identifier s'il répond ou pas au besoin d'information spécifié dans le *topic* considéré. Le document est alors qualifié de pertinent ou de non-pertinent.

De telles collections de test ont été mises en œuvre à plusieurs reprises dans des cadres d'évaluation tels que TREC et GeoCLEF. Notons qu'ils ne prennent pas en compte les trois dimensions de l'information géographique. C'est pourquoi nous proposons d'adapter leur constitution pour évaluer la RI géographique, en fournissant :

1) des *topics* couvrant tout ou partie des trois dimensions. Par exemple, un *topic* pourrait avoir pour titre « *Transhumance dans les Alpes au XIX^e siècle* » et pour nar-

ration « *Seront considérés pertinents les documents évoquant la transhumance ou les événements rattachés (quotidien du berger en estive) dans le massif des Alpes entre 1800 et 1899* » ;

2) un *corpus* traitant des trois dimensions : l'aspect thématique classiquement considéré est complété par des éléments spatiaux et temporels ;

3) des *qrels* par dimension où l'assesseur évalue l'adéquation entre chacune des trois dimensions considérées (thématique, spatiale et temporelle) et le document. Notons que la seule présence des trois dimensions dans le document ne suffit pas à déduire qu'il est pertinent pour la requête. Considérons par exemple le cas d'un document traitant du thermalisme, puis citant « *Gavarnie* » en tant que lieu de naissance du narrateur. Bien que pertinent spatialement, il ne répond pas à la requête « *thermalisme à Gavarnie* ». C'est en raison de ce type de subtilité que l'assesseur doit également évaluer l'adéquation globale entre la requête et le document.

Concernant le jugement de chaque document, l'assesseur évalue son adéquation avec chacune des trois dimensions. Cette adéquation est actuellement booléenne pour ne pas surcharger les assesseurs ; ce choix rejoint les observations de Bucher *et al.* (2005) qui soulignent que les jugements graduels par dimension sont inutilement complexes à réaliser. À partir des trois jugements booléens et du jugement global également booléen, la valeur de pertinence $v \in \{0; 1; 2; 3; 4\}$ du document est constituée. Cette valeur traduit d'une part le nombre de dimensions pertinentes et d'autre part la pertinence globale. Notons qu'aucune hypothèse n'est faite sur l'importance relative des dimensions, elles sont considérées équitablement.

4) des *ressources géographiques* nécessaires, d'une part, au géoréférencement des entités spatiales et, d'autre part, à l'interprétation des entités temporelles contenues dans le corpus.

Le protocole expérimental détaillé dans la section suivante vise à mesurer l'efficacité des SRI. Ces derniers sont évalués à partir de leur *runs* : l'ensemble des documents restitués par topic.

3.2. Protocole d'analyse comparative des SRI géographiques

La tâche évaluée est une recherche qualifiée de *ad hoc* dans TREC : le SRI répond à un besoin d'information par une liste de documents ordonnée par pertinence décroissante. L'évaluation vise à mesurer l'efficacité relative des SRI suivants :

- SRI monodimensionnel : thématique (Th), spatial (S) et temporel (Te) ;
- SRI bidimensionnels : Th+S, Th+Te et S+Te permettant de mesurer l'apport de chacune des dimensions dans l'efficacité du SRI ;
- SRI géographique combinant les trois dimensions : Th+S+Te.

Pour un *topic* donné, chaque SRI fournit une liste de couples (d, s) représentant le score s de chaque document d restitué. Classiquement, l'efficacité d'un SRI est évaluée grâce aux mesures *Average Precision (AP)* pour chaque topic et *Mean Average*

Precision (MAP) globalement. Ces dernières requièrent des *qrels* booléens (Manning *et al.*, 2008, ch. 8). Or, dans le protocole expérimental proposé, les *qrels* sont graduels afin de représenter les trois dimensions de l'information géographique. Ces deux mesures ne sont donc pas adaptées. C'est pourquoi nous recourons à la mesure de pertinence graduelle *Normalized Discounted Cumulative Gain (NDCG)* proposée par Järvelin *et al.* (2002) et utilisée notamment dans le cadre de la campagne d'évaluation TREC-9 pour la tâche Web caractérisée par des *qrels* graduels (Voorhees, 2001). Cette mesure implémente deux principes. D'une part, les documents très pertinents ($v \rightarrow 4$ dans notre cas) sont plus intéressants que les documents peu pertinents ($v \rightarrow 1$). D'autre part, un document a d'autant moins d'intérêt pour l'utilisateur que son rang est élevé dans la liste de résultats, car il est d'autant moins probable que l'utilisateur accède à ce document-là.

À l'image du protocole d'expérimentation de TREC, nous proposons deux niveaux de granularité d'évaluation d'un SRI : 1) le niveau topic en calculant *NDCG* et 2) le niveau global en calculant la moyenne arithmétique *moyNDCG* des n valeurs de *NDCG*, fournissant ainsi la mesure globale de performance du SRI.

Au niveau global, les n différences observées $\langle m_i^1 - m_j^1, \dots, m_i^n - m_j^n \rangle$ sont rapportées en pour-cent (d'amélioration ou de détérioration), où m_s^t représente la valeur de la mesure m obtenue par le système s pour le topic t . La significativité des tests statistiques calculée pour les différences observées est également rapportée : les p -valeurs de significativité sont calculées avec le test t de Student pairé (la différence est calculée entre les paires de valeurs m_i^t et m_j^t) et bilatéral (car $\forall t \in [1; n] m_i^t \not\approx m_j^t$). Bien que nécessitant théoriquement une distribution normale des données, Hull (1993) précise que ce test est en pratique robuste aux violations de cette condition. Par ailleurs, Sanderson *et al.* (2005) montrent que ce test est bien plus fiable que d'autres, tel que le test des rangs signés de Wilcoxon. Concrètement, lorsque $p < \alpha$ avec $\alpha = 0,05$ la différence entre les deux échantillons testés est qualifiée de statistiquement significative (Hull, 1993). Plus la valeur p est petite, plus la différence est significative.

4. Étude de cas : évaluation du prototype de SRI géographique PIV

Afin de valider l'hypothèse selon laquelle combiner les trois dimensions améliore la qualité des résultats, nous exploitons le cadre proposé sur la base du prototype PIV. Pour ce faire, la section 4.1 présente ce prototype. Puis, la section 4.2 détaille la mise en œuvre du cadre d'évaluation dans le but de mesurer la qualité des résultats.

4.1. Le SRI géographique PIV

Cette section décrit des éléments de conception relatifs au prototype de SRI géographique PIV. En particulier, ses trois chaînes d'indexation dédiées sont détaillées en section 4.1.1. Par la suite, la section 4.1.2 est consacrée à la description du processus d'interrogation mis en œuvre.

4.1.1. Indexation : chaînes thématique, temporelle et spatiale dédiées

Conformément aux préconisations de Clough *et al.* (2006), nous traitons indépendamment chacune des trois dimensions de l'information géographique : spatiale, temporelle et thématique. Ceci implique la construction de plusieurs index, un par dimension au moins, comme le préconise (Martins *et al.*, 2005). La recherche monodimensionnelle et la gestion des index (ajout de nouveaux documents dans le corpus) restent efficaces. Au-delà, cette approche vise la combinaison de ces index dans le cadre de recherches d'information multidimensionnelle. Elle contribue au domaine de la RI géographique telle que défini par Jones *et al.* (2006). La figure 1 décrit les trois chaînes d'indexation dédiées au traitement de documents textuels dans le prototype PIV.

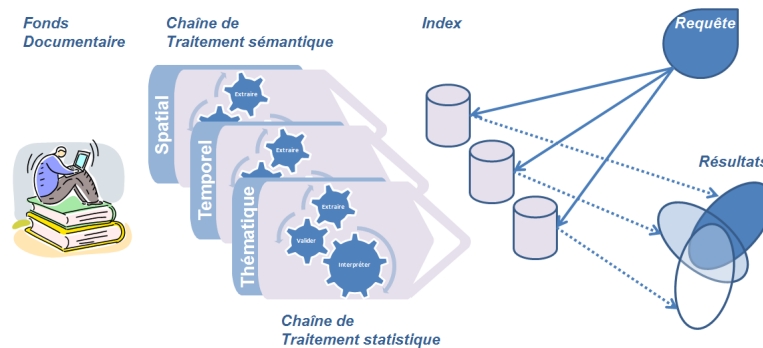


Figure 1. Les chaînes de traitement de PIV dédiées aux corpus de documents textuels.

Chaque chaîne de traitement génère un index spécialisé. Les chaînes spatiales et temporelles sont supportées par des modules de traitement automatique de la langue. Elles conduisent à l'extraction et à l'interprétation d'entités spatiales (ES) et temporelles (ET) contenues dans des documents textuels : « *le gave de Pau* » est annoté ES absolue tandis que « *au nord du gave de Pau* » est annoté ES relative – relation spatiale d'orientation (Gaio *et al.*, 2008) ; de même « *le printemps 1840* » est annoté ET absolue tandis que « *vers le printemps 1840* » est annoté ET relative – relation temporelle d'adjacence (Le Parc-Lacayrelle *et al.*, 2007). Ainsi, la création des index spatial et temporel est réalisée en trois étapes. La première étape consiste à extraire les ES et les ET à l'aide d'une chaîne de traitement syntaxico-sémantique (Gaio *et al.*, 2008). Cette chaîne est supportée par la plateforme linguastream (Bilhaut *et al.*, 2003; Widlöcher *et al.*, 2005). Elle est composée principalement d'une analyse lexicale, d'une analyse morpho-syntaxique et d'une analyse syntaxico-sémantique réalisée à l'aide d'une grammaire DCG (*Definite Clause Grammar*) dont l'objectif est d'associer un type et une sémantique aux ES et ET détectées. La deuxième étape consiste à interpréter les représentations symboliques ainsi associées à chaque ES et ET. L'interprétation est supportée par des algorithmes d'approximation qui associent des intervalles de temps aux ET (Le Parc-Lacayrelle *et al.*, 2007) et des géométries aux ES (Gaio *et al.*, 2008),

à l'aide des opérateurs spatiaux du SIG PostGIS². Enfin, la troisième étape procède à l'uniformisation des index ainsi constitués. Il s'agit d'un tuilage spatial, temporel et thématique qui permet notamment de mémoriser les fréquences d'évocation de tout ou partie de ces tuiles spatiales, temporelles et thématiques dans les textes (Palacio *et al.*, 2009; Palacio *et al.*, 2010). Cette approche nécessite des traitements complémentaires liés aux trois dimensions :

1) spatiale : permettant un carroyage régulier ou administratif (commune, canton, département...) de la zone couverte par le fonds documentaire et une projection des entités spatiales de l'index (et de leur fréquence d'apparition) dans ce carroyage ;

2) temporelle : permettant un carroyage régulier ou calendaire (jour, semaine, mois...) de la période couverte par le fonds documentaire et une projection des entités temporelles de l'index (et de leur fréquence d'apparition) dans ce carroyage ;

3) thématique : permettant un carroyage (basé sur les concepts issus d'ontologies de domaine spécifiques) des sujets couverts par le fonds documentaire et une projection des termes de l'index (et de leur fréquence d'apparition) dans ce carroyage.

L'intérêt de cette étape d'uniformisation est double. Elle permet tout d'abord de ramener des représentations spatiales, temporelles et thématiques diverses à une représentation homogène supportée par un redécoupage uniforme de l'espace, du temps et du thème. Elle permet également la mise en œuvre de stratégies de RI et de calculs de scores de pertinence éprouvés basés sur les fréquences d'apparition de ces carreaux spatiaux, temporels ou thématiques dans les unités documentaires. Ces aspects sont repris dans la section suivante qui traite de l'interrogation de ces index.

4.1.2. Interrogation : combinaison des listes de résultats

Nous avons retenu un carroyage administratif communal (resp. mensuel) pour l'indexation spatiale (resp. temporelle). À ces carroyages, nous avons appliqué des calculs de fréquence discrets et continus (proportionnels à la surface de recouvrement entre l'ES et la tuile spatiale, ou bien entre l'ET et la tuile temporelle). Pour la partie RI, nous avons préalablement expérimenté différents modèles de RI (TF, TF-IDF, OkapiBM25) que nous avons appliqués à des pondérations discrètes et continues. Les résultats de notre étude nous ont permis de retenir la formule du TF associée à une pondération continue (Palacio *et al.*, 2009; Palacio *et al.*, 2010). La chaîne de traitement thématique n'étant pas encore totalement automatisée, nous proposons d'évaluer nos travaux relatifs à l'espace et au temps et de limiter la partie thématique au système d'indexation plein-texte Terrier (Ounis *et al.*, 2005).

Ainsi, chaque SRI monodimensionnel est indépendant : il construit et interroge un index qui lui est propre. Dans notre approche, le SRI géographique proposé repose sur plusieurs SRI monodimensionnels (sources) dont les résultats sont combinés pour ne former qu'une seule liste de résultats *l*. Or, dans la littérature de RI, Fox *et al.* (1993) ont proposé à cet effet le combinateur CombMNZ. La liste combinée *l* comprend alors tous les documents distincts restitués par les SRI sources. Dans cette liste, la

2. <http://postgis.refractions.net>

similarité s d'un document d est calculée en additionnant les similarités de d issues des sources, cette somme étant pondérée par le nombre de SRI sources ayant restitué d . Ainsi, pour une requête q donnée, d sera d'autant plus pertinent dans l (c'est-à-dire classé en tête de l) qu'il a été restitué par de nombreux SRI en tête de liste (similarité s élevée entre d et q). Le comportement de CombMNZ est assimilable au principe du faisceau de preuves : des documents restitués par plusieurs SRI constituent autant d'indices renforçant la présomption de pertinence eu égard à ces documents. Pour un autre contexte combinant les dimensions thématique et sémantique, ce principe a précédemment été validé expérimentalement dans (Hubert *et al.*, 2009).

De plus, Lee (1997) a comparé les résultats de CombMNZ avec d'autres opérateurs sur des collections de test de TREC, montrant son efficacité. C'est cette qualité avérée qui nous a motivé à utiliser cet opérateur pour combiner les résultats des SRI monodimensionnels. Les similarités s qu'ils calculent étant très variables d'un SRI à l'autre – appartenant à des intervalles réels différents – nous les normalisons sur $[0; 1]$ au préalable grâce à l'équation 1 proposée dans (Lee, 1997).

$$\text{similarité_normalisée} = \frac{\text{similarité_non_normalisée} - \text{similarité_minimum}}{\text{similarité_maximum} - \text{similarité_minimum}} \quad [1]$$

Ainsi, pour une requête donnée $q = 8$, la figure 2 (a-c) illustre les résultats des trois SRI, chacun étant composé de couples (d_i, s) où d_i est un document et s est la similarité calculée entre q et d_i . Le résultat de la combinaison des SRI est représenté en figure 2 (d) où les valeurs des similarités résultant de CombMNZ sont présentées avec le détail des calculs associés, basés sur les similarités des sources (a-c) normalisées. Sur cet exemple, on observe le fait que le score d'un d_i dépend de deux facteurs. Plus un d_i est souvent restitué par les SRI, plus il obtient un score s élevé. De plus, ce dernier est d'autant plus élevé que le d_i était initialement classé en tête de liste. Le document d_4 illustre ce principe.

4.2. Mise en œuvre du cadre expérimental pour évaluer le SRI géographique PIV

Afin d'évaluer le SRI géographique PIV, nous avons mis en œuvre le cadre expérimental présenté en section 3. Nous détaillons successivement la collection de test constituée, les analyses comparatives réalisées, ainsi que les limites identifiées.

4.2.1. Constitution de la collection de test MIDR_2010

La collection de test MIDR_2010³ comprend les quatre volets identifiés en section 3.1. Premièrement, le *corpus* documentaire représente 5 645 paragraphes issus de 11 ouvrages numérisés (et traités avec une application de reconnaissance de caractères) qui proviennent du fonds patrimonial de la médiathèque. Un document d

3. Médiathèque Intercommunale à Dimension Régionale, située à Pau en Aquitaine. La collection de test est accessible sur <http://t2i.univ-pau.fr/MIDR>.

(a) SRI thématique			(b) SRI spatial			(c) SRI temporel		
q	d	s	q	d	s	q	d	s
8	d_4	14,5	8	d_8	150	8	d_8	1
8	d_3	12	8	d_1	120	8	d_4	0,7
8	d_7	8,7	8	d_4	80	8	d_9	0,5
8	d_1	0,5	8	d_9	-10	8	d_1	0,5
			8	d_2	-30	8	d_2	0,5

(d) SRI géographique résultant de la combinaison des trois résultats		
q	d	Similarité s calculée avec CombMNZ
8	d_4	$6,0333 = 3 \times \left(\frac{14,5-0,5}{14,5-0,5} + \frac{80+30}{150+30} + \frac{0,7-0,5}{1-0,5} \right)$
8	d_8	$4,0000 = 2 \times \left(\frac{150+30}{150+30} + \frac{1-0,5}{1-0,5} \right)$
8	d_1	$2,5000 = 2 \times \left(\frac{0,5-0,5}{14,5-0,5} + \frac{120+30}{150+30} \right)$
8	d_3	$0,8214 = 1 \times \left(\frac{12-0,5}{14,5-0,5} \right)$
8	d_7	$0,5857 = 1 \times \left(\frac{8,7-0,5}{14,5-0,5} \right)$
8	d_9	$0,2222 = 2 \times \left(\frac{-10+30}{150+30} + \frac{0,5-0,5}{1-0,5} \right)$
8	d_2	$0,0000 = 2 \times \left(\frac{-30+30}{150+30} + \frac{0,5-0,5}{1-0,5} \right)$

Figure 2. Principe de combinaison de résultats de recherche avec CombMNZ.

restitué à l'utilisateur par le SRI géographique est un de ces paragraphes, vu comme le meilleur point d'entrée dans l'ouvrage associé. Deuxièmement, 31 *topics* couvrant tout ou partie des trois dimensions de l'information géographique ont été constitués. Troisièmement, les *qrels* ont été obtenus en interrogeant trois SRI – un thématique basé le modèle PL2 (configuration de base de Terrier), un spatial et un temporel – avec le titre des *topics*. Quatrièmement, les ressources géographiques liées au corpus sont issues de l'Institut Géographique National (BD NYME®) et d'un *gazetteer* (un dictionnaire géographique) contributif local, ainsi que d'une base de connaissances calendaire. Pour chaque *topic*, les résultats restitués par tous les SRI ont été pris en compte pour constituer le pool. Ce dernier a été évalué, en considérant un jugement booléen par dimension ainsi qu'un jugement booléen global, ces quatre évaluations étant agrégées pour constituer un jugement graduel, comme expliqué en section 3.1.

4.2.2. Analyse comparative des résultats des SRI

Le tableau 1 présente les comparaisons effectuées entre les différents SRI et deux « baselines » thématiques (SRI de référence) identifiées dans (Perea-Ortega *et al.*, 2008) : Th^+ est une baseline forte correspondant au modèle OkapiBM25 et Th^- est une baseline faible correspondant au modèle TF-IDF. Le SRI spatial est noté S et le SRI temporel est noté Te (cf. section 3.2). Ces résultats portent sur l'analyse des performances des moteurs au regard des 31 requêtes.

Fusion de N SRI	SRI monodimensionnels				moyNDCG	Amélioration (%)	
	Th ⁻	Th ⁺	S	Te		Th ⁻	Th ⁺
1	✓				0,4726	0,0	0,1
		✓			0,4721	-0,1	0,0
			✓		0,4574	-3,2	-3,1
				✓	0,4836	2,3	2,4
2	✓	✓			0,4723	-0,1	0,0
	✓		✓		0,6004*†	27,0	27,2
	✓			✓	0,6698*†	41,7	41,9
		✓	✓		0,6005*†	27,1	27,2
			✓	✓	0,6707*†	41,9	42,1
				✓	0,7010*†	48,3	48,5
3	✓	✓	✓		0,6025*†	27,5	27,6
	✓	✓		✓	0,6672*†	41,2	41,3
	✓		✓	✓	0,7382*†	56,2	56,4
		✓	✓	✓	0,7393*†	56,4	56,6
4	✓	✓	✓	✓	0,7362*†	55,8	55,9

Tableau 1. Efficacité des SRI par rapport aux baselines thématiques. Le symbole ‘*’ (resp. †) indique une différence significative par rapport à la baseline Th⁻ (resp. Th⁺).

Globalement, il est intéressant de noter que ces deux baselines sont quasiment identiques. Contrairement aux résultats rapportés par Perea-Ortega *et al.* (2008), TF-IDF fournit de meilleurs résultats qu’OkapiBM25 dans notre cas. Cette différence peut être due au fait que la collection MIDR_2010 est décomposée en paragraphes de document de tailles homogènes (*versus* des documents entiers de tailles variables). Or, la bonne performance d’OkapiBM25 s’observe notamment sur un corpus aux documents de tailles variables, ce qui n’est pas le cas dans le présent corpus.

Concernant les moteurs monodimensionnels, on observe une performance maximale de 0,4836 pour le SRI temporel. Ces SRI sont toutefois caractérisés par une performance similaire. Par ailleurs, le gain apporté par la combinaison de dimensions hétérogènes est observable à partir de deux dimensions combinées. Notons que ce gain est statistiquement significatif au regard des deux baselines. Nous remarquons que l’alliance du spatial et du temporel apporte les meilleures performances (0,7010). Toutefois, la combinaison du thématique et du temporel offre une performance similaire (0,6707). C’est certainement dû au fait qu’une entité spatiale absolue, telle que « *Gavarnie* », est détectable par un SRI thématique. Cependant, des situations plus complexes faisant intervenir des entités spatiales relatives ne pourront être correctement traitées qu’avec un SRI spatial.

Enfin, la combinaison des trois dimensions apporte la performance maximale (0,7393). Le gain de précision correspondant à 56,6 % par rapport à la baseline forte Th⁺ valide l’hypothèse à la base de notre travail : combiner les trois dimensions de l’information géographique améliore la pertinence des résultats par rapport à la dimension thématique seule. Par contre, une analyse approfondie montre que combi-

ner les trois dimensions n'offre pas des résultats significativement supérieurs (5,5 %, $p = 0,078$) à la meilleure combinaison de deux dimensions (S+Te). Notons qu'une combinaison des trois dimensions avec les deux versions thématiques (Th⁺ et Th⁻) n'apporte pas d'amélioration supplémentaire (0,7362), le renforcement thématique masquant alors les aspects complémentaires apportés par les deux autres dimensions.

4.2.3. Limites actuelles de l'évaluation

En l'état, l'expérimentation proposée dans cet article présente au moins deux limites. Premièrement, de par ses 5 645 paragraphes totalisant 3,7 Mo, la collection de test MIDR_2010 est très peu volumineuse en comparaison des collections TREC. Deuxièmement, nous avons réalisé les analyses à partir de 31 topics. Or, cette valeur représente seulement six topics de plus que le nombre minimum de topics à considérer pour pouvoir réaliser des analyses statistiques valides. Nous continuons l'effort de jugement manuel des documents permettant d'obtenir davantage de topics à analyser.

Malgré ces limites, le cadre d'évaluation présenté dans cet article est applicable aux diverses propositions issues du domaine de la RI géographique. La section suivante présente de façon succincte un sous-ensemble représentatif de ces travaux.

5. Approches de la littérature en RI géographique

Des travaux connexes au projet PIV sont menés dans la communauté internationale; nous pouvons notamment citer les six projets suivants. Le projet DIGMAP (Manguinhas *et al.*, 2009) signifiant "*Discovering our Past World with Digitised Maps*" est un système spécialement axé sur le matériel historique et la valorisation du patrimoine culturel et scientifique. Le projet GéoSem (Bilhaut *et al.*, 2003) est un système dédié au traitement sémantique pour l'information géographique (textes, cartes, graphiques). Le projet GIPSY (Woodruff *et al.*, 1994) propose une méthode d'indexation de documents textuels basée sur l'agrégation des géo-références correspondant aux entités spatiales trouvées dans le texte. L'idée est d'utiliser cette agrégation pour retrouver la zone géographique la plus représentative, qui servira à indexer le document. Le projet GRID (Valcartier, 2006) signifiant "*Geospatial Retrieval of Indexed Documents*" est un système dédié à la recherche d'information textuelle combinant une recherche par mots-clés et une recherche par zone d'intérêt via une interface cartographique. Le projet SPIRIT (Vaid *et al.*, 2005) signifiant "*Spatially-Aware Information Retrieval on the Internet*" est un système dédié à la recherche de pages web faisant référence à des lieux ou zones géographiques spécifiés dans une requête. Enfin, le projet STEWARD (Lieberman *et al.*, 2007) signifiant "*Spatio-Textual Extraction on the Web Aiding Retrieval of Documents*" est un système dédié à l'extraction, l'interrogation et la visualisation de références à des zones géographiques dans des textes non structurés. Comme indiqué dans le tableau 2, ces systèmes d'indexation et de RI géographique traitent en priorité la composante spatiale (entités ESA et ESR, cf. section 4.1.1). Seul le système SPIRIT produit des index indépendants et des index multidimensionnels. Pour les autres systèmes, les approches de combinaison envisagées

sont généralement limitées au choix d'une dimension initiale puis, à l'application des autres dimensions sur le sous-ensemble de documents obtenu (priorité aux termes puis au spatial, pour STEWARD par exemple). Il s'agit donc d'une démarche séquentielle de filtrage et non d'une vraie combinaison.

Système	Entités spatiales		Nature des index			
	ESA	ESR	spatial	temporel	thématique	multidim.
DIGMAP	✓		✓	✓	✓	
GEOSEM	✓	✓	✓	✓	✓	
GIPSY	✓		✓			
GRID	✓		✓		✓	
PIV	✓	✓	✓	✓	✓	
SPIRIT	✓		✓		✓	✓
STEWARD	✓		✓		✓	

Tableau 2. Comparaison de projets et prototypes dédiés à la RI spatiale.

En complément de la comparaison fonctionnelle des systèmes présentée dans le tableau 2, il serait opportun de comparer leur efficacité. Notre cadre d'évaluation permettra cette comparaison dans les deux scénarios suivants. Les contributeurs évaluent eux-mêmes leur système à partir du cadre proposé, ou bien ils mettent leur système à disposition afin que nous puissions les évaluer (ce qui n'est pas le cas actuellement).

6. Conclusion et perspectives

Dans cet article, nous avons considéré les SRI géographiques exploitant les dimensions spatiale, temporelle et thématique. Or, les moteurs de recherche usuels s'avèrent limités dans de tels contextes. Aussi, notre contribution se décline en deux volets : la proposition d'un cadre d'évaluation ainsi que la validation de notre hypothèse initiale (combinaison des trois dimensions améliore la qualité des résultats de recherche). L'application de ce cadre sur une collection de test appropriée a mis en exergue une amélioration de 56,5 % qui est, de plus, statistiquement significative. Ces bons résultats apportent une validation empirique de nos propositions expérimentées avec le prototype PIV (Gaio *et al.*, 2008). Par ailleurs, ne se limitant pas à ces trois seules dimensions, ce cadre d'évaluation peut aussi intégrer d'autres dimensions telles que la confiance en l'information et sa fraîcheur (Costa Pereira *et al.*, 2009).

En complément d'expérimentations sur une collection de test plus volumineuse, nous envisageons désormais d'expérimenter et de proposer de nouvelles approches de combinaison. Nous nous intéresserons plus particulièrement aux approches de combinaison par contraintes (prenant en compte les notions d'exigences et de préférences) selon une approche linéaire (Farah *et al.*, 2008), ou selon une approche basée sur la logique floue avec les opérateurs OWA (Yager, 1988; Boughanem *et al.*, 2007), ou encore, sur d'autres propositions comme l'intégrale de Choquet (Labreuche *et al.*, 2003).

Ayant montré dans cet article la faisabilité d'une évaluation de SRI géographique, nous envisageons également de proposer cette tâche pour une campagne de type GeocLEF, son originalité étant de traiter des documents français sur les trois dimensions : thématique, spatiale et temporelle. La collection de test MIDR_2010 constituée est d'ores et déjà disponible sur le site web du projet PIV⁴.

Remerciements

Les auteurs remercient vivement Valérie Bougeant, Jean-Paul Cabanac, Christian Chevalier, Benoît Chirle, Agnès Itier, Anaïs Lefeuvre, Nicolas Noullet et Camille Raymond pour leur contribution bénévole en qualité d'assesseurs.

7. Bibliographie

- Bilhaut F., Charnois T., Enjalbert P., Mathet Y., "Geographic reference analysis for geographic document querying", *HLT-NAACL'03: Proceedings of the workshop on Analysis of geographic references*, ACL, Morristown, NJ, USA, p. 55–62, 2003.
- Boughanem M., Loiseau Y., Prade H., "Refining Aggregation Functions for Improving Document Ranking in Information Retrieval", *SUM'07: Proceedings of the 1st international conference on Scalable Uncertainty Management*, Springer-Verlag, Berlin, Heidelberg, p. 255–267, 2007.
- Bucher B., Clough P., Joho H., Purves R., Syed A. K., "Geographic IR Systems: Requirements and Evaluation", *ICC'05: Proceedings of the 22nd International Cartographic Conference*, Global Congressos, 2005. CDROM.
- Buckley C., Voorhees E. M., "Evaluating Evaluation Measure Stability", *SIGIR'00: Proceedings of the 23rd international ACM SIGIR conference*, ACM, New York, NY, USA, p. 33–40, 2000.
- Clough P., Joho H., Purves R., "Judging the Spatial Relevance of Documents for GIR", *ECIR'06: Proceedings of the 28th European Conference on IR Research*, vol. 3936 of LNCS, Springer, p. 548–552, 2006.
- Costa Pereira C., Dragoni M., Pasi G., "Multidimensional Relevance: A New Aggregation Criterion", *ECIR'09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, Springer-Verlag, Berlin, Heidelberg, p. 264–275, 2009.
- Farah M., Vanderpooten D., "An outranking approach for information retrieval", *Inf. Retr.*, vol. 11, n° 4, p. 315–334, 2008.
- Fox E. A., Shaw J. A., "Combination of Multiple Searches", in D. K. Harman (ed.), *TREC-I: Proceedings of the First Text REtrieval Conference*, NIST, Gaithersburg, MD, USA, p. 243–252, February, 1993.
- Gaio M., *Traitements de l'information géographique : Représentations et structures*, Habilitation à diriger des recherches, Université de Caen, 2001.
- Gaio M., Sallaberry C., Etcheverry P., Marquesuzaa C., Lesbegueries J., "A global process to access documents' contents from a geographical point of view", *J. Vis. Lang. Comput.*, vol. 19, n° 1, p. 3–23, 2008.

4. <http://t2i.univ-pau.fr/MIDR>

- Gan Q., Attenberg J., Markowetz A., Suel T., "Analysis of geographic queries in a search engine log", *LocWeb'08: Proceedings of the first international workshop on Location and the web*, ACM, New York, NY, USA, p. 49–56, 2008.
- Gey F. C., Larson R. R., Sanderson M., Joho H., Clough P., Petras V., "GeoCLEF'05: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview", *CLEF'05: Proceedings of the 6th workshop on Cross-Language Evaluation Forum*, vol. 4022 of *LNCS*, Springer, p. 908–919, 2006.
- Gospodnetić O., Hatcher E., *Lucene in Action*, Manning Publications, 2005.
- Harman D. K., "The TREC Test Collections", in Voorhees *et al.* (2005), chapter 2, p. 21–53, 2005.
- Hubert G., Mothe J., "An adaptable search engine for multimodal information retrieval", *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, n° 8, p. 1625–1634, 2009.
- Hull D., "Using Statistical Testing in the Evaluation of Retrieval Experiments", *SIGIR'93: Proceedings of the 16th annual international ACM SIGIR conference*, ACM Press, New York, NY, USA, p. 329–338, 1993.
- Järvelin K., Kekäläinen J., "Cumulated gain-based evaluation of IR techniques", *ACM Trans. Inf. Syst.*, vol. 20, n° 4, p. 422–446, 2002.
- Jones C. B., Purves R., "GIR'05 2005 ACM workshop on geographical information retrieval", *SIGIR Forum*, vol. 40, n° 1, p. 34–37, 2006.
- Jones R., Zhang W. V., Rey B., Jhala P., Stipp E., "Geographic intention and modification in web search", *Int. J. Geogr. Inf. Sci.*, vol. 22, n° 3, p. 229–246, 2008.
- Kanhubua N., Nørvåg K., "Improving Temporal Language Models for Determining Time of Non-timestamped Documents", *ECDL'08: Proceedings of the 12th European conference on Research and Advanced Technology for Digital Libraries*, Springer-Verlag, Berlin, Heidelberg, p. 358–370, 2008.
- Labreuche C., Grabisch M., "The Choquet integral for the aggregation of interval scales in multicriteria decision making", *Fuzzy Sets Syst.*, vol. 137, n° 1, p. 11–26, 2003.
- Le Parc-Lacayrelle A., Gaio M., Sallaberry C., "La composante temps dans l'information géographique textuelle", *Document Numérique*, vol. 10, n° 2, p. 129–148, 2007.
- Lee J. H., "Analyses of Multiple Evidence Combination", *SIGIR'97: Proceedings of the 20th annual international ACM SIGIR conference*, ACM Press, New York, NY, USA, p. 267–276, 1997.
- Lieberman M. D., Samet H., Sankaranarayanan J., Sperling J., "STEWART: Architecture of a Spatio-Textual Search Engine", *GIS'07: Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*, ACM, New York, NY, USA, p. 1–8, 2007.
- Manguinhas H., Martins B., Borbinha J., Siabato W., "The DIGMAP Geo-Temporal Web Gazetteer Service", *e-Perimetron: Int. Web J. Sci. Technol. Affined Hist. Cartogr. Maps*, vol. 4, n° 1, p. 9–24, 2009.
- Manning C. D., Raghavan P., Schütze H., *Introduction to Information Retrieval*, Cambridge University Press, July, 2008.
- Martins B., Silva M. J., Andrade L., "Indexing and ranking in Geo-IR systems", *GIR '05: Proceedings of the 2005 workshop on Geographic information retrieval*, ACM, New York, NY, USA, p. 31–34, 2005.

- Ogilvie P., Callan J. P., “Experiments Using the Lemur Toolkit”, *TREC’01: Proceedings of the 9th Text REtrieval Conference*, NIST, Gaithersburg, MD, USA, February, 2001.
- Ounis I., Amati G., Plachouras V., He B., Macdonald C., Johnson D., “Terrier Information Retrieval Platform”, *ECIR’05: Proceedings of the 27th European Conference on IR Research*, vol. 3408 of *LNCS*, Springer, p. 517-519, 2005.
- Palacio D., Sallaberry C., Gaio M., “Normalizing Spatial Information to Better Combine Criteria in Geographical Information Retrieval”, *ECIR-GIIW’09: Proceeding of the international workshop on Geographic Information on the Internet*, p. 37–48, 2009.
- Palacio D., Sallaberry C., Gaio M., “Normalizing Spatial Information to Improve Geographical Information Indexing and Retrieval in Digital Libraries”, *ISGIS’10: Proceedings of the Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science proceedings*, 2010. to appear.
- Perea-Ortega J. M., García-Cumbreras M. A., García-Vega M., Ureña-López L. A., “Comparing Several Textual Information Retrieval Systems for the Geographical Information Retrieval Task”, *NLDB’08: Proceedings of the 13th international conference on Natural Language and Information Systems*, Springer-Verlag, Berlin, Heidelberg, p. 142–147, 2008.
- Peters C., “Introduction”, *CLEF’01: Proceedings of the 1st Workshop Cross-Language Information Retrieval and Evaluation*, vol. 2069 of *LNCS*, Springer, p. 1–6, 2001.
- Sallaberry C., Baziz M., Lesbegueries J., Gaio M., “Towards an IE and IR System Dealing with Spatial Information in Digital Libraries – Evaluation Case Study”, *ICEIS’07: Proceedings of the 9th International Conference on Enterprise Information Systems*, p. 190–197, 2007.
- Sanderson M., Kohler J., “Analyzing Geographic Queries”, *SIGIR-GIR’04: Proceedings of the Workshop on Geographic Information Retrieval at SIGIR*, 2004.
- Sanderson M., Zobel J., “Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability”, *SIGIR’05: Proceedings of the 28th annual international ACM SIGIR conference*, ACM, New York, NY, USA, p. 162–169, 2005.
- Vaid S., Jones C. B., Joho H., Sanderson M., “Spatio-textual Indexing for Geographical Search on the Web”, *SSTD’05: Proceedings of the 9th international Symposium on Spatial and Temporal Databases*, vol. 3633 of *LNCS*, Springer, p. 218–235, 2005.
- Valcartier, GRID – geospatial retrieval of indexed document, Rapport technique, R&D pour la défense, Canada, 2006.
- Verhagen M., Gaizauskas R., Schilder F., Hepple M., Moszkowicz J., Pustejovsky J., “The TempEval challenge: identifying temporal relations in text”, *Lang. Resour. Eval.*, vol. 43, n° 2, p. 161–179, 2009.
- Voorhees E. M., “Evaluation by Highly Relevant Documents”, *SIGIR’01: Proceedings of the 24th annual international ACM SIGIR conference*, ACM, New York, NY, USA, p. 74–82, 2001.
- Voorhees E. M., Harman D. K., *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press, Cambridge, MA, USA, 2005.
- Widlöcher A., Bilhaut F., “La plate-forme LinguaStream : un outil d’exploration linguistique sur corpus”, *TALN’05: Actes de la 12^e Conférence sur le Traitement Automatique du Langage Naturel*, 2005.
- Woodruff A. G., Plaunt C., “GIPSY: automated geographic indexing of text documents”, *J. Am. Soc. Inf. Sci.*, vol. 45, n° 9, p. 645–655, 1994.
- Yager R. R., “On ordered weighted averaging aggregation operators in multicriteria decision-making”, *IEEE Trans. Syst. Man Cybern.*, vol. 18, n° 1, p. 183–190, 1988.