# Indexation multi-critères et différentes approches de combinaison

**Damien Palacio**

*Laboratoire LIUPPA*
*Université de Pau et des Pays de l'Adour*
*UFR Sciences et Techniques*
*Avenue de l'Université – B.P. 1155*
*64013 PAU Cedex*

*damien.palacio@univ-pau.fr*

RÉSUMÉ. *Ce papier s'inscrit dans la continuité de travaux sur l'indexation et la recherche d'information menés au LIUPPA sur des critères géographiques. L'information géographique a trois composantes : le spatial, le temporel et le thématique. Notre équipe a déjà travaillé sur le spatial et le temporel de façon indépendante. Aujourd'hui nous cherchons à combiner ces différentes composantes. Pour cela nous proposons d'utiliser une approche statistique, réservée habituellement à l'analyse plein-texte d'un document, pour le spatial et le temporel. Cette approche sera qualifiée de « carroyage ». Toutefois nous cherchons aussi à extraire les liens sémantiques existants entre les informations géographiques. Pour cela nous étudions un autre type de combinaison, qualifiée d'« approche par motifs ».*

ABSTRACT. *This paper is following works on indexation and information retrieval made at the LIUPPA on geographics criteria. Geographic information has three components : spatial, temporal and thematic. Our team has already worked on spatial and temporal independantly. Today we want to combine these differents components. We propose to use a stastical aproach, usually reserved to full-text analyse of a document, for spatial and temporal. This aproach will be called "grid". However, we want also to extract semantics links existing between geographics informations. We study an other combinaison type, called "approach by pattern".*

MOTS-CLÉS : *indexation, recherche d'information, spatial, temporel, appariement, combinaison, cadres, grille.*

KEYWORDS: *indexation, information retrieval, spatial, temporel, matching, combination, frame, grid.*

## 1. Contexte

La quantité de documents numérisés croît de façon très importante ces dernières années, grâce à l'amélioration des techniques de numérisation, à la réduction des coûts et aux avantages apportés : gain de place (serveurs ou DVD), pérennisation des documents (réplication à divers endroits), nouvelles facilités d'accès (internet). Ainsi la recherche d'information fine devient un point critique. Contrairement à beaucoup de travaux s'intéressant aux pages web, nous travaillons sur des livres. Le fonds documentaire utilisé est un ensemble de documents textuels patrimoniaux peu structurés (l'OCR[1] étant simple, la structure logique a été perdue) fournis par la médiathèque de Pau (MIDR).

Deux observations, la première qu'une part non négligeable de requêtes d'utilisateurs a trait à la géographie (Sanderson *et al.*, 2004), et la deuxième que de nombreux fonds documentaires sont constitués de documents patrimoniaux territorialisés, démontrent l'intérêt de travailler sur l'information géographique.
Une information géographique peut être considérée comme une molécule à trois composantes : le temps, l'espace et le thème/phénomène (Lesbegueries *et al.*, 2006). Une requête géographique utilisant une recherche « plein-texte » [2] pourra ne pas donner de résultats pertinents. Par exemple, si nous faisons une recherche « les villes au sud de Pau », n'apparaîtront pas dans les résultats les documents contenant « Jurançon », or c'est une ville au sud de Pau.

Ce papier propose une réflexion sur les différentes manières de combiner l'utilisation d'index temporel, spatial et thématique pour une recherche d'information géographique efficace.

Dans un contexte de RI, nous cherchons à rendre la combinaison de ces trois types d'index homogène et efficace. Pour cela, nous proposons de retraiter ces index spatiaux et temporels selon une approche statistique vectorielle similaire à celles proposées aujourd'hui pour l'indexation de termes dans les documents. Nous allons plus particulièrement étudier des méthodes de pondération de cellules/tuiles correspondant au découpage d'une zone géographique et de pondération d'intervalles de temps correspondant au découpage d'une période temporelle.

Les approche statistiques font généralement leurs preuves dans l'évaluation de la pertinence mais restent confinées aux données fournies, si ces dernières sont locales, il sera impossible de déterminer leurs portées [3]. Nous nous intéressons donc à établir l'existence de relations sémantiques entre entités spatiales et temporelles en étudiant leurs portées. Pour cela nous cherchons à redécouper les documents, auparavant en syntagmes [4], par des motifs constitués d'ensemble de syntagmes.

---

1. optical character recognition ou reconnaissance optique de caractères en français
2. Une recherche plein-texte ou dite full-text ne cherche que les termes donnés
3. La portée d'une entité est la portion/étendue de texte sur laquelle porte cette entité
4. Un syntagme est une unité syntaxique située entre le mot et la phrase, cela correspond donc à des expressions. Par exemple « au sud de Pau »

## 2. Des chaînes de traitement dédiées

Les travaux envisagés dans ce papier s'appuient sur des résultats déjà obtenus au sein de notre équipe de recherche. Dans (Sallaberry *et al.,* 2007) et (Le-Parc-Lacayrelle *et al.,* A paraître) les auteurs ont travaillé sur la composante spatiale et sur la composante temporelle. Deux chaînes de traitements sémantiques ont été réalisées permettant d'extraire et d'indexer automatiquement des informations spatiales et temporelles. Ces chaînes sont indépendantes.

### 2.1. *Indexation spatiale*

Une première chaîne de traitement a été mise au point (via un prototype nommé PIV) [5] réalisant à la fois l'extraction et l'indexation d'entités spatiales (ES). Cette chaîne comporte deux parties distinctes : une première réalisant un traitement sémantique (analyse du texte puis balisage d'une liste d'entités spatiales candidates), et une deuxième basée sur un système d'information géographique (SIG) afin de valider tout ou partie des entités spatiales détectées et de les géoréférencer [6] (Sallaberry *et al.,* 2007).

Deux types d'entités spatiales ont été définis : les entités spatiales absolues (ESA) et les entités spatiales relatives (ESR). Chaque entité spatiale est représentée par un polygone. Les ESA sont des entités spatiales de référence telle qu'une commune, un pic ou encore une forêt. Par exemple, la ville de Pau est une ESA ; ses coordonnées et le polygone qui la représente sont connus. Les ESR sont des entités spatiales basées sur des ES. Chaque ESR est définie par une relation avec une ou plusieurs ES et sa représentation est calculée selon sa sémantique (type de la relations, et ES). Un certain nombre de relations ont été défini dans le modèle : orientation, adjacence, inclusion, distance, figure géométrique et intersection (Lesbegueries, 2007). Par exemple l'ESR « au Sud de Pau » désigne une relation d'orientation par rapport à l'ESA Pau.

### 2.2. *Indexation temporelle*

De même que pour le spatial, une chaîne de traitement dédiée à l'extraction et à l'indexation des informations temporelles a été réalisée. Cette chaîne comporte également deux parties : une première marque toutes les entités temporelles et une deuxième leur associe une représentation sémantique plus abstraite afin d'approximer la période couverte par chaque entité (Le-Parc-Lacayrelle *et al.,* A paraître).

Ici aussi il y a deux types d'entités temporelles (ET) : les entités temporelles absolues (ETA) et les entités temporelles relatives (ETR). Une entité temporelle est représentée par un intervalle de temps (le jour étant le grain le plus fin).

---

5. PIV pour Pyrénées Itinéraires Virtuels : prototype de plateforme de recherche d'information pour le projet du même nom

6. Géoréférencer signifie référencer une donnée via des coordonnées géographiques.

Les ETA sont des entités temporelles de références, telles que des dates. Les ETR sont des entités temporelles basées sur des ET. Les ETR sont définies par une relation avec une ou plusieurs ET : adjacence, orientation, inclusion, intervalle, énumération (Le-Parc-Lacayrelle *et al.,* A paraître). Par exemple « au début de l'année 1950 » est une ETR composée d'une relation d'inclusion de type « début » avec l'ETA « année 1950 ». Elle sera représentée par une période allant du 1er janvier 1950 au 30 avril 1950 (nous retenons le premier tiers).

### 2.3. *Recherche d'information*

Les deux chaînes de traitements sémantiques précédentes permettent de produire deux index indépendants : un spatial et un temporel. A cela, peut être rajouté un index thématique obtenu via des approches statistiques classiques, qualifiées aussi de full-text, qui extraient et indexent les termes.
Une fois ces index produits, il est alors possible de s'intéresser à la recherche d'informations. Les requêtes sont en texte libre et sont traitées par les chaînes de traitements présentées auparavant (la chaîne spatiale pour une requête spatiale, et la chaîne temporelle pour la requête temporelle). En ce qui concerne l'appariement, pour chacune des deux approches un algorithme a été proposé, dans (Sallaberry *et al.,* 2007) pour le spatial et dans (Le-Parc-Lacayrelle *et al.,* A paraître) pour le temporel. La recherche d'informations ne s'effectuant que sur un critère est donc complètement cloisonnée.

## 3. Combiner pour une RI Géographique

### 3.1. *Travaux Actuels*

Trois index ont donc été élaborés de manière autonome : un spatial, un temporel et un thématique. Or l'intérêt n'est pas de faire des requêtes spécifiques à chaque composante, mais au contraire de pouvoir faire de véritables requêtes géographiques (par exemple « instruments de musique dans les environs de Laruns au XIXème siècle »). Le problème est donc de déterminer des méthodes adéquates pour combiner ces différents index.

Devons nous garder des index bien distincts ou bien, au contraire, les fusionner pour n'en garder qu'un. Dans (Martins *et al.,* 2005) et (Vaid *et al.,* 2005) les auteurs ont évalué les différentes possibilités et concluent qu'il est nettement plus avantageux de conserver les index de chaque composante séparés. Cela permet notamment de pouvoir faire des recherches spécifiques à un critère sans perte d'efficacité, ou de ne mettre à jour qu'un type d'index sans modifier le reste.
Dans (Martins *et al.,* 2005), les auteurs proposent différentes formules pour réaliser une combinaison simple des critères spatiaux et thématiques : combinaisons linéaires, similarité maximum . . . Mais dans cet article, ainsi que dans (Vaid *et al.,* 2005), le score de pertinence de chaque critère n'est pas calculé de la même manière. Une comparaison directe de ces critères ne peut donc pas être réalisée.

Pour simplifier la représentation des informations spatiales, les auteurs de (Martins *et al.*, 2005) ont proposés de réaliser un découpage d'une zone géographique en « quadrillage » avec notamment la technique des C-Squares[7] définis par (Rees, 2003). Cette méthode de découpage est particulièrement intéressante, car elle permet un partitionnement récursif sans limite, c'est à dire qu'on peut décomposer une zone autant de fois que nous le souhaitons et obtenir une bonne précision, et en plus elle utilise la latitude et la longitude pour localiser les cases (« squares »).

Concernant nos index une première approche a été proposée par (Sallaberry *et al.*, 2007). La requête est découpée en deux sous-requêtes, l'une spatiale et l'autre thématique contenant tout ce qui n'a pas été reconnu comme spatial. Chaque sous-requête donne un ensemble de résultats. Pour combiner, nous faisons l'intersection entre ces deux ensembles, donc seuls les documents présents dans chacun des deux ensembles est pertinent. Au final nous obtenons de meilleurs résultats qu'en utilisant une seule approche : la précision est de 70% contre 48% pour l'approche classique seule et 15% pour l'approche spatiale seule. Mais le rappel est faible (quatre réponses alors que respectivement 724 pour le sous-ensemble spatial et 233 pour le sous-ensemble thématique) à cause de la combinaison trop « restrictive ». De plus les indices de pertinence ne sont pas homogènes et l'approche reste très locale.

### 3.2. *Nos propositions*

Deux propositions sont abordées ici. La première porte sur une approche statistique vectorielle mettant en oeuvre une méthode dite de carroyage. Elle permet notamment d'évaluer la pertinence de tout type d'entité de manière homogène. La deuxième est une approche sémantique, nommée approche par motifs. Elle permet d'extraire les relations entre les entités et de déduire leurs portées. Ces deux approches sont complémentaires, l'approche de carroyage pouvant s'appliquer aussi bien sur les index des chaînes de traitements actuelles ou sur les index qui seraient générés par l'approche par motifs.

Tout d'abord, nous avons défini une première hypothèse : pour combiner de manière homogène les différents critères il faut que le calcul de pertinence soit similaire. Nous proposons alors d'appliquer une approche statistique vectorielle, déjà utilisée pour le thématique, au spatial et au temporel. Pour la rendre applicable nous envisageons de mettre en oeuvre une méthode de segmentation de l'espace d'information. Pour le spatial cela consiste à découper une zone géographique (un planisphère si nous voulons couvrir toutes les zones possibles) en cellules/tuiles, telles que les C-Squares proposés dans (Rees, 2003). Pour le temporel, il s'agit de découper une période temporelle calendaire en intervalle de temps.

Une fois les découpages effectués, il faut pondérer chaque tuile/intervalle. Or il existe une formule très utilisée pour les approches statistiques, le tf.idf, se basant sur

7. Concise Spatial Query And Representation System

la fréquence du terme dans un document et dans la base documentaire. Nous cherchons donc à réaliser un calcul similaire pour nos tuiles/intervalles. Dans ce cas TF correspond à Tile Frequency (Fréquence des tuiles), et l'idf est appliqué de la même façon.
Sur le schéma de la figure 1 sont représentées deux ESA (la ville de Pau en B1-B2-C2 et la ville de Jurançon en A2-A3-B2-B3) et une ESR ("le sud de Pau" représenté par une boîte englobante), illustrant un résultat possible de cette approche. A titre d'exemple la pondération des ES contenues dans les tuiles est binaire (voir formule 1), c'est à dire que la surface d'intersection n'est pas pris en compte. La pondération d'une tuile est la somme des poids de toutes les ES qui l'intersectent. Comme illustré sur la figure 1, la tuile B2 a un poids de 3 car on a trois entités. Bien évidemment, l'objectif sera ici de trouver une fonction permettant de pondérer de manière pertinente les tuiles, par exemple en tenant compte du pourcentage d'occupation d'une tuile par les ES (voir formule 2).
En ce qui concerne le temporel, l'idée serait de réaliser des formules du même type pour des intervalles de temps.
Cette approche se basant sur des index existants (ceux produits par nos chaînes de traitements), cela implique la création de nouveaux index pour contenir ces segments produits et leurs pertinences.

$$TF_{Tuile} = \Sigma_{ES=1}^{n} W_{ES} \quad avec\ W_{ES} = \begin{cases} 1\ si\ ES \cap Tuile \neq \oslash \\ 0\ sinon \end{cases} \qquad [1]$$

avec Tuile la tuile étudiée, W le poids d'une ES, n le nombre d'ES

$$W_{ES} = \frac{S_{ES}}{S_{Tuile}} \quad avec \left| \begin{array}{l} S_{ES} \text{ la superficie de la tuile occupée par l'ES} \\ S_{Tuile} \text{ la superficie de la tuile} \end{array} \right. \qquad [2]$$

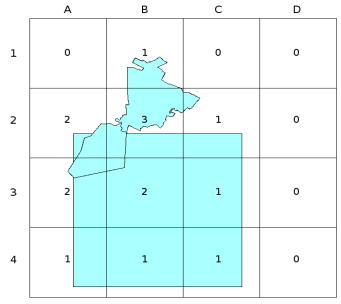


**Figure 1.** *Exemple de carroyage*

Malgré le fait que cette approche de carroyage semble très prometteuse, elle reste néanmoins dépendante des index. Par conséquent si ceux ci proviennent d'une exploi-

tation locale de l'information alors il est impossible de connaître le poids informatif (ou portée) d'une expression retenue comme index.

La seconde proposition s'appuie donc sur l'hypothèse qu'il existe des liens sémantiques entre les entités. Certains de ces liens peuvent être modélisés par des motifs (itinéraire, comparaison, énumération, ...). Parmi ces motifs, certains peuvent être mis en évidence grâce à la théorie des Cadres (Charolles, 1997). Différents travaux ont permis de rendre opératoire cette théorie et ont validé son intérêt par des expérimentations sur des corpus (Laignelet, 2004) (Bilhaut, 2006). Nous nous proposons d'étudier dans une problématique de RI, l'apport de la théorie des cadres sur la constitution de motifs. Nous proposons alors de réaliser de nouveaux index contenant cette fois non plus des entités mais des motifs.
Chaque cadre a une portée, et elle se détermine notamment grâce aux entités spatiales et temporelles qui le composent. Un cadre contiendra un introducteur de cadre (IC), c'est à dire une entité placée en début de phrase, qui déterminera le type du cadre (spatial ou temporel). Toutes les entités contenues dans un cadre ont de très fortes chances d'être sémantiquement liées avec l'introducteur de ce cadre. Cela correspond donc à sa portée
Voici un exemple (figure 2) présentant deux cadres (délimités par des crochets) introduits par des entités temporelles (respectivement « printemps 1787 » et « En 1799 »). Dans le premier cadre, il y a une ETR, « entre l'hiver et l'été », qui est effectivement liée à l'IC Temporel, confirmant la période (printemps). Dans un premier temps l'objectif est de réaliser des cadres temporels, et des cadres spatiaux, puisque travaillant indépendamment sur chacun des index générés, puis de les combiner.

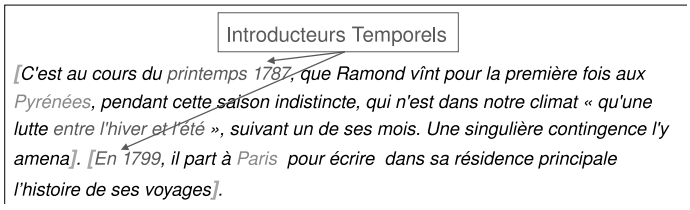


**Figure 2.** *Exemple de cadres*

Indexer des documents sur des critères géographiques peut apporter une nette amélioration dans la recherche d'information ciblée. Cependant cela implique de combiner efficacement les différents critères. L'approche de carroyage permet d'évaluer leurs pertinences de manière homogène, mais n'apporte aucune information supplémentaire à celles fournies par les index au niveau sémantique. L'approche par motifs, notamment celle des cadres, permet d'extraire certaines relations sémantiques entre entités et de déterminer leurs portées. Toutefois l'approche par carroyage peut aussi s'appliquer sur les index qui seraient générés par l'approche par motifs et pondérer ces motifs au lieu des entités.

## 4. Conclusion

Nous avons parlé dans ce papier de la manière de combiner des index géographiques (spatial, temporel et thématique). Pour cela deux approches ont été proposées, à la fois indépendantes et complémentaires. L'approche de carroyage (statistique) consiste à découper des informations géographiques et à pondérer les cellules/intervalles produites. Sur cette approche, il s'agit de déterminer une formule pertinente pour pondérer les cellules/intervalles en fonction des entités qui les intersectent (importance, pourcentage d'occupation, ...). L'approche par motifs (sémantique) consiste à extraire des liens entre entités, et d'évaluer leurs portées. Ici il s'agit donc d'extraire et indexer des cadres du discours, et obtenir des relations entre entités.

## 5. Bibliographie

Bilhaut F., Analyse automatique de structures thématiques discursives, PhD thesis, Université de Caen, 2006.

Charolles M., « L'encadrement du discours - Univers, champs, domaines et espaces », *Cahiers de Recherche Linguistique*, vol. 6, p. 1-73, 1997.

Laignelet M., *« Les titres et les cadres de discours temporels »*, Master's thesis, Université de Toulouse 2 - Le Mirail, 2004.

Le-Parc-Lacayrelle A., Gaio M., Sallaberry C., « La composante temps dans l"information géographique textuelle », *Document Numérique*, A paraître.

Lesbegueries J., Plate-forme pour l'indexation spatiale multi-niveaux d'un corpus territorialisé, PhD thesis, Université de Pau et des Pays de l'Adour, 2007.

Lesbegueries J., Gaio M., Loustau P., « Geographical information access for non-structured data. », *Proceedings of the 2006 ACM Symposium on Applied Computing (SAC), Dijon, France*, p. 83-89, 2006.

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.

Rees T., « "C-Squares", a New Spatial Indexing System and its Applicability to the Description of Oceanographic Datasets », *Oceanography*, vol. 16, p. 11-19, 2003.

Sallaberry C., Baziz M., Lesbegueries J., Gaio M., « Une approche d'extraction et de recherche d'information spatiale dans les documents textuels - Evaluation », *Actes de la Conférence en Recherche d'Informations et Applications, CORIA 2007*, 2007.

Salton G., Buckley C., « Term-weighting approaches in automatic text retrieval », *Information Processing & Management*, vol. 24, n° 5, p. 513-523, 1988.

Sanderson M., Kohler J., « Analyzing Geographic Queries », *Proceedings of the Workshop on Geographic Information Retrieval, SIGIR*, 2004.

Vaid S., Jones C. B., Joho H., Sanderson M., « Spatio-textual Indexing for Geographical Search on the Web. », *in* , C. B. Medeiros, , M. J. Egenhofer, , E. Bertino (eds), *SSTD*, vol. 3633 of *Lecture Notes in Computer Science*, Springer, p. 218-235, 2005.