Indexation spatiale et temporelle baséee sur un principe de ”tuilage”: contribution à la recherche d'information géographique dans des documents textuels faiblement structurés

Christian Sallaberry, Damien Palacio, Mauro Gaio

Indexation spatiale et temporelle basée sur un principe de « tuilage »

Contribution à la recherche d'information géographique dans des documents textuels faiblement structurés

Christian Sallaberry — Damien Palacio — Mauro Gaio

LIUPPA, Université de Pau, BP 1155, 64000 PAU (FRANCE)
nom.prenom@univ-pau.fr

RÉSUMÉ. *La plupart des moteurs de recherche nécessitent, pour fonctionner, une indexation préalable des documents. Certaines de ces approches sont limitées compte tenu de contextes particuliers ou de la forme particulière de l'information recherchée. Notre contribution porte sur la construction d'index adaptés à la facette spatiale et temporelle spécifique au contexte de l'information géographique tout en permettant une compatibilité avec les outils de recherche génériques. Ce travail présente une stratégie générique d'indexation basée sur le principe du « tuilage ». Elle s'applique aussi bien sur la composante spatiale que temporelle mais peut également être utilisée sur la composante thématique. Nous évaluons ensuite l'apport de cette approche à la recherche d'information géographique.*

ABSTRACT. *Most of search engines process users' information needs by retrieving documents from pre-built term-based indexes. Such approaches are limited regarding particular contexts or specific retrieval criteria. Our contribution concerns geographical information retrieval (GIR) and proposes to exploit both spatial and temporal facets to extend classical thematic engines in order to parse unstructured textual documents. This work proposes a general tile-based indexing strategy that can be applied to spatial, temporal or thematic information. It evaluates the contribution of such a process to geographic information retrieval approches.*

MOTS-CLÉS : *Indexation, Recherche d'Information Géographique*

KEYWORDS: *Indexing, Geographical Information Retrieval*

1. Introduction

Nous reprenons la théorie selon laquelle l'information géographique comporte trois facettes : spatiale, temporelle et thématique (Longley *et al.,* 2005). L'extrait de texte *« Les villes et les châteaux fortifiés dans le bassin aquitain au* XIII^{e} *siècle »* illustre bien ce triptyque. Toutefois, les textes analysés ne décrivent pas toujours ces trois facettes tel que dans l'exemple précédent ; seules une ou deux facettes sont parfois explicitées.

Le principal objectif de PIV (Pyrénées Itinéraires Virtuels) est de rechercher les informations pertinentes dans des livres (Gaio *et al.,* 2008). Nous proposons de compléter les approches de recherche d'information (RI) classiques en intégrant les facettes géographiques de l'information dans un processus de recherche. Ainsi, cette contribution vise l'uniformisation de l'information géographique par la création d'index composés de tuiles spatiales et temporelles. Cette étape permet, d'une part, la mise en ?uvre d'une stratégie de pondération de tuiles éprouvée en RI. D'autre part, elle est un préalable nécessaire à la RI géographique combinant des critères spatiaux, temporels et thématiques, telle que nous l'envisageons.

Les projets SPIRIT (Vaid *et al.,* 2005), STEWARD (Lieberman *et al.,* 2007), CITER (Pfoser *et al.,* 2009) et DIGMAP (Manguinhas *et al.,* 2009) font référence à des systèmes d'indexation et de recherche d'information géographique qui traitent la facette spatiale en priorité. Comme le souligne R.R. Larson (Larson, 2009), les systèmes de RI (SRI) spatiaux classiques associent un ou des focus spatiaux à un document à partir d'un ensemble d'entités spatiales (ESs). Une géométrie est calculée et associée à chaque ES. Ainsi, la pertinence spatiale d'un document est calculée sur la base de la surface de recouvrement entre la géométrie correspondant à la requête et celles correspondant aux ESs décrivant le document. Les projets CITER et DIGMAP complètent ces systèmes dédiés aux informations spatiales par des SRI similaires dédiés aux informations temporelles. Enfin, tous proposent des SRI dédiés à la facette thématique. Concernant la combinaison en phase d'interrogation, les systèmes CITER et SPIRIT adoptent une approche de « filtrage parallèle ». Ils interrogent simultanément et séparément chaque facette, puis combinent les différentes listes de résultats en réalisant leur intersection. D'autres systèmes, tel que STEWARD, mettent en œuvre une approche de « filtrage séquentiel » qui consiste à interroger le corpus sur une facette puis à appliquer les autres facettes sur le sous-ensemble de documents obtenu. D'autres encore, tel que DIGMAP, proposent d'utiliser des approches de combinaison linéaire (moyenne arithmétique) sans toutefois uniformiser les critères, ce qui permet de pondérer les résultats mais peut biaiser le score global.

Les travaux relatifs à la RI définissent l'uniformisation comme un processus de lemmatisation de termes afin de les regrouper et de leur associer des poids (Spärck Jones, 1972). L'un des modèles les plus populaires développé en RI est le modèle vectoriel de Salton (Salton *et al.,* 1983). Il définit une représentation d'une collection de documents sous la forme d'une matrice de document-par-termes (D. Manning *et*

al., 2008). Ce modèle supporte la recherche et le classement de documents par calcul de mesures de similarité entre un document et une requête.

Par conséquent, nous proposons d'étendre cette notion de matrice de document-par-termes et de mettre en oeuvre une matrice de document-par-tuiles afin d'uniformiser les index spatiaux et temporels. Cette approche permet de construire des matrices de document-par-tuiles dans lesquelles l'élément (i, j) d'une matrice décrit la fréquence d'occurrence de la tuile i dans le document D_j.

L'article est structuré comme suit. La section 2 est consacrée à la description formelle de l'indexation par tuilage spatial et temporel. La section 3 présente succinctement l'exploitation de tels index dans un processus de RIG. Enfin, nous concluons en section 4.

2. Indexation par tuilage spatial et temporel

Nous utilisons les chaines de traitements du SRI géographique PIV pour produire un premier niveau d'index spécialisés (Gaio *et al.,* 2008) (Le Parc-Lacayrelle *et al.,* 2007). Nous uniformisons ensuite ces index pour produire un second niveau d'index à base de tuiles.

2.1. *Uniformisation par tuilage : un exemple sur des tuiles spatiales ou temporelles*

Prenons l'exemple d'un texte contenant la phrase « Je suis passé près de Bayonne au début du mois de janvier 2001 ». Les chaines de traitement PIV vont analyser ce texte et produire un premier niveau d'index associant une géométrie à l'entité spatiale (ES) « près de Bayonne » (voir SF5, Figure 1(a)) et un intervalle de temps à l'entité calendaire (EC) « au début du mois de janvier 2001 » (voir CF9, Figure 1(b)).

L'uniformisation par tuilage va ensuite se dérouler en deux étapes : le choix de la segmentation pour chaque facette puis, l'application de ce tuilage sur les informations de chaque index de premier niveau. Ainsi, l'uniformisation spatiale nécessite un tuilage régulier ou administratif (commune, canton, département) de la zone couverte par le fonds documentaire et un calcul d'intersection des ESs de l'index (et de leur fréquence d'apparition) et de ce tuilage. De même, l'uniformisation temporelle nécessite un tuilage régulier ou calendaire (jour, semaine, mois) de la période couverte par le fonds documentaire et un calcul d'intersection des ECs de l'index (et de leur fréquence d'apparition) et de ce tuilage.

La Figure 1(a) illustre le tuilage administratif communal de la zone. L'ES « près de Bayonne » (SF5) intersecte sept tuiles (communes autour de Bayonne). De même, la Figure 1(b) illustre le tuilage calendaire mensuel de la période. L'EC « au début du mois de janvier 2001 » (CF9) est incluse dans la tuile T9 (mois de janvier 2001).

L'extrait d'index présenté dans la Table 1 illustre l'application du tuilage présenté sur la Figure 1(b) ; les tuiles T1, T2,. . . , Tn et les fréquences d'évocation correspon-

id_t	id_{sf} list	fréquence binaire cumulée	fréquence proportionnelle cumulée
T1	[]	0	0
T2	[CF6]	1	0.21
T3	[CF2 ;CF6]	2	1.08
T4	[CF3 ;CF6]	2	1.98
...			

Tableau 1. *Index à base de tuilage temporel*

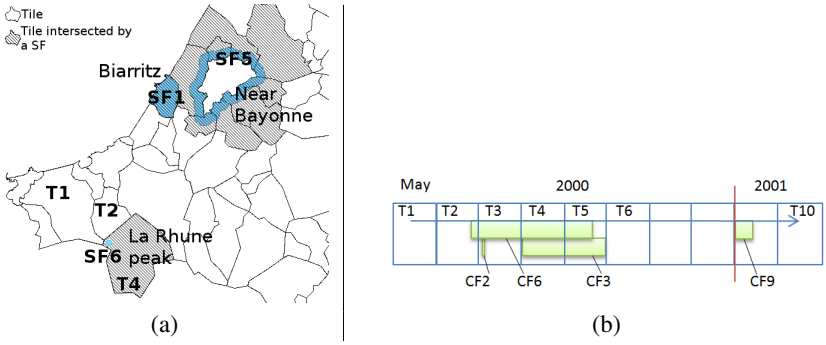

Figure 1. *Application du tuilage : (a) Entité spatiales sur un tuilage administratif communal, (b) Entités temporelles sur un tuilage calendaire mensuel.*

dantes. Par exemple, CF6 intersecte quatre tuiles (T2, T3, T4, T5) ; la fréquence binaire relative à chacune de ces tuiles sera augmentée de 1. La tuile T3, quant à elle, est en intersection avec deux entités temporelles (CF6 et CF2) ; par conséquent, elle aura un poids standard de 2. La fréquence proportionnelle tient compte de la surface de recouvrement entre la représentation d'une entité et une tuile. Elle tient également compte de la granularité de l'entité par rapport à celle de la tuile. Par exemple, un tel calcul déterminera une fréquence proportionnelle de 1,08 pour l'application de CF2 et CF6 sur la tuile T3.

Ces stratégies d'indexation nous permettent de proposer plusieurs index à base de tuilage spatial et calendaire : des index spatiaux dont les segmentation sont de niveau quartier, ville, canton, département, région, pays, ainsi que des index calendaires dont les segmentations sont de niveau heure, jour, semaine, mois, saison, année, siècle. Dans un contexte RIG, ces index back-office permettent à un moteur de recherche de parcourir l'index le plus adapté au grain spatial et calendaire de la requête.

2.2. *Uniformisation par tuilage : formalisation de la démarche*

Nous proposons d'appliquer les méthodes de RI usuelles telle que l'uniformisation des représentations (lemmatisation, troncature) et les modèles de pondération aux index spécifiques tels que ceux dédiés aux informations spatiales ou temporelles. L'approche de tuilage consiste à appliquer aux représentations initiales un tuilage à autant de dimensions que comportent ces représentations. Cette approche est une forme de discrétisation : par exemple, elle associe un objet temporel du domaine calendaire à une tuile temporelle correspondant à une segmentation particulière du même espace.

De manière plus formelle, au domaine[1] O inclus dans l'espace R^n [2] correspond un domaine T inclus dans l'espace R^n. Le domaine O est constitué d'un ensemble d'objets $O_1, \ldots, O_p$ et le domaine T est constitué de l'union de m sous-espaces (les tuiles). Pour chaque sous-espace de T en intersection avec 1 ou plusieurs objets de O on retient le nombre d'intersections (N_{T_i}).

$$
\begin{aligned}
O \subseteq \mathbb{R}^n \longrightarrow T \subseteq \mathbb{R}^n & \\
O = \{O_1, O_2, O_3, \ldots, O_p\} & \\
T = \bigcup_{i=1}^{m} T_i & \qquad [1] \\
N_{T_i} = \mid T_i \mid T_i \cap O_j \neq \varnothing \quad \forall j = 1, \ldots, p \mid & \\
\text{avec } \mid x \mid \text{ la cardinalité de } x &
\end{aligned}
$$

Cette approche utilise donc un index spécifique existant et génère un nouvel index uniformisé. Ainsi à un ensemble de représentations à une dimension, nous appliquons un tuilage à une dimension, à un ensemble de représentations à deux dimensions, un tuilage à deux dimensions, et ainsi de suite jusqu'à n dimensions. Par exemple (Figure 1), l'index temporel spécifique référence des représentations calendaires visualisables sur la ligne de temps (une dimension). L'index uniformisé correspondant comporte des tuiles (semaines, mois, ...) matérialisées sur la même ligne de temps. De même l'index spatial spécifique référence des représentations géométriques visualisables sur un fonds cartographique (deux dimensions). L'index uniformisé correspondant comporte des tuiles (communes, départements, ...) matérialisées sur la même carte.

Après avoir choisi le tuilage, nous pouvons pondérer les tuiles en utilisant les approches basées sur les fréquences d'apparition. Pour calculer la fréquence d'une tuile nous proposons deux approches discrètes. La fréquence binaire consiste à compter le

1. Un domaine est un ensemble fini ou infini de valeurs. On le représente par une liste d'éléments ou bien une condition nécessaire et suffisante d'appartenance, le domaine des booléens : {0,1}, le domaine des doigts de la main : {pouce, index, majeur, annulaire, auriculaire}, le domaine calendaire

2. Ce sur-ensemble modélise les espaces de dimension 1, 2 ou plus

nombre de représentations initiales de l'information (objets) qui intersectent la tuile (un objet pouvant intersecter plusieurs tuiles). La fréquence proportionnelle se base sur le ratio de recouvrement entre un objet et une tuile, la fréquence est ainsi incrémentée d'une valeur comprise entre 0 et 1. Le tableau 2 détaille les formules de ces deux types de fréquences.

Fréquence binaire	$freq(T_i) = \sum_{j=1}^{p} freq(T_i, O_j)$
Fréquence proportionnelle	$freqP(T_i) = \sum_{j=1}^{p} freq(T_i, O_j) * \frac{Surf(T_i, O_j)}{Surf(T_i)} * \frac{1}{NbTuiles(O_j)}$

Tableau 2. *Formules de fréquence ($freq(T_i, O_j)$: fréquence de l'objet O_j dans la tuile T_i (nombre d'intersection), $Surf(T_i, O_j)$: surface de l'objet O_j dans la tuile T_i, $Surf(T_i)$: surface de la tuile T_i, $NbTuiles(O_j)$: nombre de tuiles intersectées par l'objet O_j)*

Cette approche d'indexation basée sur les tuiles permet l'utilisation de modèles usuels pour calculer des scores de pertinences basés sur les fréquences d'occurrence des tuiles dans les documents et des formules éprouvées telles que TF.IDF ou OkapiBM25 (D. Manning *et al.,* 2008).

3. Application à la RIG

3.1. *RI mono-facette et RIG multifacette*

Nous avons mis au point trois SRI : le SRI spatial PIV_v2, le SRI temporel PIV_v2 et le SRI multifacette PIV_v3 combinant les deux autres SRI et le SRI thématique Terrier (Ounis *et al.,* 2005). Le SRI spatial PIV_v2 et le SRI temporel PIV_v2 génèrent et exploitent, respectivement, plusieurs index de granularité différente. L'index le mieux adapté au grain de la requête est utilisé en phase d'interrogation. Nous avons testé les deux approches de pondérations discrètes (Table 1) associées à plusieurs formules très répandues en RI pour évaluer ces deux SRI monofacettes : TF, TF.IDF et OkapiBM25. Le SRI multifacette PIV_v3 repose sur les SRI mono-facettes PIV_v2 spatial, PIV_v2 temporels et le SRI thématique Terrier dont les résultats sont combinés pour ne former qu'une seule liste de résultats l : nous utilisons le combinateur CombMNZ défini par Fox et Shaw (Fox *et al.,* 1993) pour réaliser une combinaison linéaire. La liste combinée l comprend tous les documents d distincts restitués par les SRI sources. Pour une requête q donnée, d sera d'autant plus pertinent dans l qu'il a été restitué par de nombreux SRI en tête de liste.

3.2. *Premiers résultats*

Le *corpus* documentaire étudié représente 5 645 paragraphes issus de 11 ouvrages qui proviennent du fonds patrimonial de la médiathèque. Ceci correspond notamment

à 10 968 entités spatiales et 1 702 entités temporelles. Un document d restitué à l'usager par le SRI géographique est un de ces paragraphes, vu comme le meilleur point d'entrée dans l'ouvrage associé. Un ensemble de 31 *topics* couvrant tout ou partie des trois facettes de l'information géographique a été constitué (Palacio *et al.,* 2010).

L'expérimentation du SRI spatial PIV_v2 montre que la formule TF associée à la pondération proportionnelle (TFp) donnent les meilleurs résultats. Le tuilage administratif communal est, par défaut, celui qui sera le plus adapté à ce jeu de requêtes. L'expérimentation comparant les résultats de PIV_v1 à PIV_v2 montre une amélioration de la pertinence des résultats de 13% qui est, de plus, statistiquement significative.

L'expérimentation du SRI temporel PIV_v2 montre que la formule TF associée à la pondération proportionnelle (TFp) donnent également les meilleurs résultats. Le tuilage calendaire mensuel est, par défaut, celui qui sera le plus adapté à ce jeu de requêtes. La perte de précision engendrée par l'uniformisation des index n'est que partiellement compensée. Le nombre d'ECs présentes dans chaque paragraphe est généralement inférieur ou égal à un. L'exploitation de la fréquence d'évocation des tuiles dans le texte n'a donc pas ici d'influence pour l'amélioration des résultats. En effet, l'expérimentation comparant les résultats de PIV_v1 à PIV_v2 montre une stabilité de la pertinence (-0,5%).

Enfin, l'expérimentation du SRI multifacette PIV_v3, le comparant au SRI thématique Terrier montre une amélioration de la pertinence des résultats de 66,5% qui est, de plus, statistiquement significative (Palacio *et al.,* 2010). Ce résultat valide notre hypothèse initiale selon laquelle combiner les trois facettes géographiques améliore la qualité des résultats de recherche d'information.

4. Conclusion

L'intérêt de cette démarche d'uniformisation d'index par tuilage 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 tuiles spatiales, temporelles ou thématiques dans les unités documentaires. Une telle uniformisation conduisant au regroupement d'entités dans des tuiles, entraîne une perte de précision (Chrisman, 1990). Toutefois, elle est compensée par l'introduction de calculs de fréquences ainsi que la production et l'utilisation d'index de grains différents. Nous travaillons désormais sur la mise au point d'opérateurs de combinaison de résultats issus de différents SRI mono-facettes. Nous souhaitons notamment associer des degrés de préférence et d'exigence aux différents critères de recherche exprimés par l'utilisateur. Ces critères seront à la charge du SRI PIV_v3.

5. Bibliographie

Chrisman N. R., « Deficiencies of sheets and tiles : building sheetless databases », *Int. J. Geogr. Inf. Sci.*, vol. 4, n° 2, p. 157-167, 1990.

D. Manning C., Raghavan P., Schütze H., *Introduction to Information Retrieval*, Cambridge University Press, New York, 2008.

Fox E. A., Shaw J. A., « Combination of Multiple Searches », *in* , D. K. Harman (ed.), *TREC-1 : Proceedings of the First Text REtrieval Conference*, NIST, Gaithersburg, MD, USA, p. 243-252, February, 1993.

Gaio M., Sallaberry C., Etcheverry P., Marquesuzaa C., Lesbegueries J., « A global process to access documents' contents from a geographical point of view », *Journal of Visual Languages And Computing*, vol. 19, n° 1, p. 3-23, 2008.

Larson R. R., « Geographic Information Retrieval and Digital Libraries », *ECDL'09 : Proceedings of the 13th European Conference on Digital Libraries*, vol. 5714 of *LNCS*, Springer, p. 461-464, 2009.

Le Parc-Lacayrelle A., Gaio M., Sallaberry C., « La composante temps dans l'information géographique textuelle », *Revue Document Numérique*, vol. 10, n° 2, p. 129-148, 2007.

Lieberman M. D., Samet H., Sankaranarayanan J., Sperling J., « STEWARD : 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.

Longley P. A., Goodchild M. F., Maguire D. J., Rhind D., *Geographic Information Systems and Science*, John Wiley & Sons, 2005.

Manguinhas H., Martins B., Borbinha J., Siabato W., « The DIGMAP geo-temporal Web gazetteer service », *e-Perimetron : International Web journal on sciences and technologies affined to history of cartography and maps*, vol. 4(1), p. 9-24, 2009.

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., Cabanac G., Sallaberry C., Hubert G., « Measuring Effectiveness of Geographic IR Systems in Digital Libraries : Evaluation Framework and Case Study », *in* , M. Lalmas, , J. Jose, , A. Rauber, , F. Sebastiani, , I. Frommholz (eds), *ECDL'10 : Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries*, vol. 6273 of *LNCS*, Springer, p. 340-351, September, 2010.

Pfoser D., Efentakis A., Hadzilacos T., Karagiorgou S., Vasiliou G., « Providing Universal Access to History Textbooks : A Modified GIS Case. », *in* , J. D. Carswell, , A. S. Fotheringham, , G. McArdle (eds), *W2GIS*, vol. 5886 of *Lecture Notes in Computer Science*, Springer, p. 87-102, 2009.

Salton G., McGill M. J., *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

Spärck Jones K., « A statistical interpretation of term specificity and its application in retrieval », *Journal of Documentation*, vol. 28, n° 1, p. 11-21, 1972.

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