Combinaison de critères par contraintes pour la Recherche d'Information Géographique

Damien Palacio

ÉCOLE DOCTORALE DES SCIENCES EXACTES ET DE LEURS APPLICATIONS ED 211

THÈSE

pour l'obtention du

Doctorat de l'Université de Pau et des Pays de l'Adour

(spécialité informatique)

présentée par

Damien PALACIO

Combinaison de critères par contraintes pour la Recherche d'Information Géographique

soutenue publiquement le 26 novembre 2010

Composition du jury

Président :	Florence SEDÈS	IRIT, Université Paul Sabatier, Toulouse
Rapporteurs :	Bénédicte BUCHER	IGN, Paris
	Éric GAUSSIER	LIG, Université Joseph Fourier, Grenoble 1
	Gabriella PASI	DiSCo, Università degli Studi di Milano - Bicocca
Examinateur :	Guillaume CABANAC	IRIT, Université Paul Sabatier, Toulouse
Directeurs :	Mauro GAIO	LIUPPA, Université de Pau et des Pays de l'Adour
	Christian SALLABERRY	LIUPPA, Université de Pau et des Pays de l'Adour

Laboratoire d'Informatique de l'Université de Pau et des Pays de l'Adour — EA 3000

LIUPPA

Mis en page avec la classe thloria.

Remerciements

Je suis vraiment ravi d'être arrivé à mener à bien cette thèse démarrée il y a maintenant trois ans. Je tiens à remercier toutes les personnes qui ont contribué de près ou de loin à cette réussite.

Je remercie Bénédicte Bucher, Éric Gaussier et Gabriella Pasi d'avoir accepté de rapporter ce mémoire, ainsi que Florence Sèdes et Guillaume Cabanac d'en être les examinateurs. Je suis très honoré d'avoir un jury si renommé !

Merci à Mauro Gaio d'avoir accepté de m'encadrer durant cette thèse et pour ses conseils avisés. Merci également à Christian Sallaberry, co-encadrant, pour sa très grande disponibilité et son aide précieuse à toutes les occasions. Je suis ravi d'avoir travaillé à vos côtés pendant ces trois ans.

Je souhaite aussi remercier nos collègues toulousains, Guillaume Cabanac et Gilles Hubert, avec qui nous avons entamé une collaboration suite à la conférence ECIR et qui a été très enrichissante et fructueuse (plusieurs articles, dont le meilleur papier d'ECDL'10). J'espère que l'année qui arrive permettra de continuer cette collaboration !

Je tiens à remercier la Communauté d'Agglomération de Pau Pyrénées d'avoir financé ce travail de thèse. J'espère que le projet d'industrialisation de nos prototypes verra le jour et permettra à la Médiathèque Intercommunale à Dimension Régionale de Pau de disposer d'un moteur de recherche plus adapté à leurs collections.

Ce fut un grand plaisir de travailler au sein de l'équipe DESI, devenue T2I en 2009, ainsi que dans le laboratoire LIUPPA et dans les locaux du département informatique de l'université. Mais aussi de partager les repas tous les midis et faire quelques sorties avec vous (CongDuc, Annig, Laurent, les deux Eric, Sophie, Bruno, Nicolas, ...). Merci plus particulièrement à Christophe pour sa compagnie quotidienne ;)

Merci à tous les anciens doctorants pour leur aide et leurs conseils : les deux Julien, Pierre, Christine et Cyril. Je souhaite aussi encourager tous les doctorants encore au laboratoire : Thanh Vu, Van Tien, Minh Duc, Nour, Éric, Natacha, Youssef, Julien, John et Camille.

Je souhaite aussi remercier tous mes amis qui m'ont supporté et encouragé pendant ce travail. Merci à François pour toutes ces choses qu'il a pu obtenir et dont il m'a fait profiter pour me distraire ;) Merci particulièrement à Patxi et Émilie d'avoir accepté de relire ce manuscrit pour corriger les (trop) nombreuses fautes restantes !!

Pour finir je remercie vivement toute ma famille de m'avoir soutenu tout au long de ce travail, d'avoir accepté mes visites moins fréquentes ou (trop) courtes et d'avoir toujours été là en cas de besoin. Merci à ma grand-mère pour ses guides très utiles lors de mes déplacements en conférence. Bon courage à mes deux frères (Sébastien et Mathieu) et à ma sœur (Mélanie) pour trouver leur voie. Enfin merci à mes parents pour tout ce qu'ils ont fait pour que je puisse en arriver là et pour leur totale confiance.

Je dédie cette thèse à ma famille.

Table des matières

Table des figures

Liste des tableaux

Première partie

Introduction : la recherche d'information géographique dans des fonds documentaires textuels

Chapitre 1

Contexte

Sommaire

1.1 Introduction

Ce manuscrit présente mes travaux de thèse financés par la Communauté d'Agglomération de Pau Pyrénées [1] et réalisés dans le Laboratoire d'Informatique de l'Université de Pau et des Pays de l'Adour (LIUPPA) [2], plus particulièrement au sein de l'équipe Document Électronique, Sémantique et Interaction (DESI) [3] devenue depuis fin 2009 l'équipe Traitement, Interaction, Information (T2I). Cette thèse s'inscrit dans la continuité des travaux de Julien Lesbegueries [Les07] ; elle vise l'accès à l'information par le contenu des documents.

Ces travaux ont été réalisés sur une collection de livres numérisés et fournis par la Médiathèque Intercommunale à Dimension Régionale (MIDR). Ce sont notamment des livres de type récits de voyages. Le tableau 1.1 présente des extraits d'un de ces livres. Néanmoins les approches proposées ici pourraient s'appliquer à d'autres types de corpus contenant des informations géographiques. Les tableaux 1.2 et 1.3 présentent deux exemples de requêtes effectuées par des utilisateurs. Les exemples présentés dans ces trois tableaux serviront de support d'illustration aux différentes discussions menées tout au long de ce manuscrit.

1. http://www.agglo-pau.fr
2. http://liuppa.univ-pau.fr
3. http://liuppa.univ-pau.fr/DESI/

Dans ce chapitre, nous allons nous intéresser au contexte de la thèse, c'est à dire aux objectifs, problématiques, hypothèses ainsi qu'aux contributions visées.

1.2 Objectif : améliorer la recherche d'information géographique en combinant des Systèmes de RI existants

Aujourd'hui la Recherche d'Information (RI) est essentiellement focalisée sur le Web. En effet, sur Internet plus de 200 millions de sites web [4] sont recensés. Les moteurs de recherche (Google, Bing, Exalead...) proposent d'aider les utilisateurs à trouver ce qui les intéresse dans cette masse d'informations.

Une étude réalisée sur les recherches scolaires a révélé que les trois catégories principales « de critères de recherche » sont : bibliographie (personnes), chronologie (périodes) et géographie (lieux) [MMBS09]. Plusieurs études montrent une part non négligeable d'informations géographiques dans les requêtes des utilisateurs : pour les moteurs Excite [SK04], AOL [GAMS08] et Yahoo [JZR+08] cette proportion varie entre 12,7% et 18,6%. Néanmoins, les moteurs de recherche usuels ne permettent pas de prendre en compte la particularité de certains types d'information, tels que le spatial ou le temporel. En effet, ils se limitent à la recherche de termes que l'utilisateur fournit dans sa requête. Si nous souhaitons trouver des documents relatant des événements associés au sud de Pau, le moteur ne va chercher que « sud » et « Pau ». Or un document évoquant « Jurançon », qui est une commune limitrophe à celle de Pau, et située à son sud, devrait aussi être retourné. De même pour le temporel, si nous souhaitons trouver des documents décrivant des événements relatifs au XX^{e} siècle, le moteur de recherche ne devrait pas seulement retourner les documents contenant « XX^{e} siècle » mais aussi ceux contenant « 1901 »,« 1902 »,...

Les sites Web ne sont pas la seule source sur laquelle porte la recherche d'information. En cette ère du tout numérique, la numérisation des documents papiers progresse en quantité et en qualité. Google par exemple, à travers son service Google Books [5], numérise massivement des livres et magazines qui sont ensuite mis à la disposition du grand public. De plus, avec l'essor des livres électroniques (appelés aussi « liseuses » ou encore *e-books*), tels que le Kindle d'Amazon ou le Reader de Sony, ainsi que des tablettes PC, tels que certains EEE d'Asus ou l'Ipad d'Apple, les versions électroniques des documents sont de plus en plus plébiscitées. Cela permet de transporter et visualiser des centaines voire des milliers de documents sur une simple carte mémoire.

La numérisation s'est longtemps limitée à la création de simples versions électroniques, c'est à dire une image par page de livre, ce qui empêche la moindre recherche et réduit les interactions possibles avec ces versions électroniques. Google Books soumet les versions numérisées à des logiciels de reconnaissance de caractères et donc propose les textes contenus dans ces livres. Il est par conséquent, possible de rechercher des mots dans les livres ainsi numérisés.

4. http://news.netcraft.com/archives/2010/07/16/july-2010-web-server-survey-16.html
5. http://books.google.fr/books

— Paragraphe 443 (d1) —
Pendant que Russell courait le monde, une autre étoile de la pléiade, Charles Packe, apparaissait à Gavarnie, cette même année 1858. [...] A Gavarnie, il y a moins à découvrir qu'ailleurs et le Grand Cirque, d'après lui, avait depuis longtemps perdu le prestige et le charme de l'inconnu.

— Paragraphe 446 (d2) —
Donc, dès son retour, fin de 1861, avec Laurent Passet, guide de Gavarnie, il va faire l'ascension du Vignemale, sa première.

— Paragraphe 461 (d3) —
[...] Russell décide superbement : le Vignemale, près de Gavarnie. Le Vignemale, le plus haut point où l'on puisse atteindre par territoire français. Et lorsque, en 1880, Russell fait sa cinquième ascension du Vignemale, c'est pour déterminer le point précis où il aménagera une grotte.

— Paragraphe 469 (d4) —
Le 8 août 1903, Henri Russell accomplit sa trente-troisième ascension et redescend avec tristesse à Gavarnie, abandonnant pour la dernière fois son glacier, ses grottes et cette cime qui était son idole. Mais son souvenir plane toujours sur cette belle montagne et sa silhouette s'évoque comme celle du roi conquérant, possesseur du roc, et poète du Vignemale.

— Paragraphe 518 (d5) —
Cette musique improvisée me remet en mémoire l'histoire du compositeur Musard qui eut vers 1840 son heure de célébrité. Il fit plusieurs voyages aux Pyrénées ; les montagnes l'inspiraient, disait-il, dans ses compositions musicales.[...]

— Paragraphe 592 (d6) —
A ce propos, l'autre soir, au refuge, nous avions le plaisir de causer avec de savants camarades — de vrais montagnards ceux-là — qui connaissent la montagne encore mieux que moi, puisqu'ils l'étudient sous tous ses aspects : sur terre, sous terre et au fond des lacs. Nous discutions sur le mot : alpiniste, employé aux quatre coins du monde pour désigner les sportsmen qui « font de la montagne ». On se sert dans les Pyrénées avec raison du mot pyrénéiste ; mais cette expression est restée strictement régionale. La raison en est que la renommée des Alpes et des ascensions alpines a été consacrée avant celle des Pyrénées, et surtout parce que, ce vocable d'origine latine a été répandu dans toute la Gaule par les armées romaines pour désigner les sommets qui leur rappelaient les hautes montagnes bornant l'ancienne Italie. La dénomination d' « Alpes » a donc été appliquée à l'époque romaine à toute région de montagnes en dehors même des Alpes proprement dites. C'est ainsi que le mot « alpage » est également utilisé un peu partout dans le sens des pâturages ou d'herbages dans la montagne, même dans les Pyrénées, bien qu'ici le mot vulgaire, pour cette désignation, soit celui de « port »

TABLE 1.1 – Extraits du livre : « Excursions autour du Vignemale dans les hautes vallées de Cauterets, de Gavarnie et du Rio Aran en Aragon » [Mei87]

« les risques accidentels en montagne si possible liés à des balades ou randonnées »

TABLE 1.2 – Exemple de requête multicritère thématique

« documents sur les montagnes des Pyrénées entre 1800 et 1900 mais pas sur Gavarnie et si possible sans rapport avec les ascensions »

TABLE 1.3 – Exemple de requête multicritère géographique

De plus en plus d'organismes tels que des médiathèques ou musées se sont lancés dans des campagnes de numérisations et d'océrisation de leurs collections. Le but est donc de permettre aux utilisateurs d'effectuer des recherches depuis n'importe où (médiathèque, domicile, téléphone, ...) grâce à une interface Web, sur tous les ouvrages de leur collection. Cela permet notamment de consulter des œuvres rares ou trop abimées pour être accessibles physiquement par tous.

Nous nous plaçons dans ce contexte de recherche d'information appliquée à des corpus de documents patrimoniaux numérisés composés de journaux, lithographies, romans, récits de voyages, ... Dans le cadre de ce travail de thèse, nous nous limitons aux récits de voyages qui sont de longs documents (plusieurs centaines de pages) et qui contiennent de nombreuses évocations spatiales et temporelles (notamment sur les Pyrénées aux XVIII-XIX^{e} siècles). Il faut noter que ces documents, fournis par la MIDR, ont été océrisés avec perte de la structure logique. Seules les ruptures de ligne ont été conservées et nous les avons considérées comme des marques de fin de paragraphe. De par la longueur de ces documents et étant donnée l'absence de leur structure, le point d'entrée choisi est le paragraphe. Ainsi, lorsqu'un utilisateur effectue une recherche, le moteur lui retourne l'ensemble des paragraphes pertinents provenant des documents du corpus. Néanmoins, pour chaque information extraite, un lien est conservé vers l'expression, le paragraphe et le document. Ainsi il est possible d'envisager des scénarios de navigation dans l'ensemble du document à partir d'un paragraphe.

Concernant les usagers, nous distinguons plusieurs catégories potentiellement intéressées par une recherche proposant des critères géographiques :

- les érudits, par exemple des historiens, souhaitant retrouver des informations précises sur un lieu ou une date.
- les archivistes pour, par exemple, améliorer les annotations des documents.
- les enseignants et leurs élèves pour, par exemple, générer l'itinéraire décrit dans un livre de type récit de voyage.
- les touristes pour, par exemple, déterminer quelles sont les activités, monuments ou autres, accessibles dans un lieu donné (« gorges au sud de Laruns », « résurgences autour de Pau », ...).

– n'importe quel utilisateur souhaitant chercher des informations avec des critères spatiaux ou temporels.

L'utilisation de traitements conduisant à des index précis et adaptés à chaque type d'information (spatiale, temporelle et thématique) permet de répondre aux différents besoins des utilisateurs. Notre objectif est ainsi d'améliorer la recherche d'information géographique en combinant les résultats obtenus par des traitements spatiaux et temporels dédiés et des stratégies classiques de recherche d'information généralement utilisées pour des critères thématiques. Il est donc nécessaire de déterminer la méthode la plus adéquate pour combiner des telles informations.

1.3 Problématique : comment combiner des critères hétérogènes de RI ?

L'hétérogénéité des données contenues dans certains documents (par exemple multimédias) nécessite leur décomposition en plusieurs critères. Par exemple, pour une vidéo, elle sera décomposée en un certain nombre d'images et une bande sonore (pouvant être convertie en texte s'il s'agit de discours). De même, l'hétérogénéité des données représentant certaines informations nécessite leur décomposition en plusieurs critères. Par exemple, selon [Use96, Gai01] l'information géographique peut être décomposée en trois facettes : le spatial, le temporel et le thématique.

Nous avons choisi de traiter chacune de ces facettes [6] spécifiquement et de manière indépendante, comme préconisé dans de nombreux travaux en Recherche d'Information Géographique (RIG) tels que [CJP06, MSA05]. Nous avons donc un système de RIG dédié conduisant à des index précis (contrairement à GeoNames qui propose des index moins précis, notamment de part la nature ponctuelle des représentations) et des méthodes de calcul adaptées pour chacune des facettes. Si nous souhaitons traiter des requêtes géographiques portant sur différents critères [7] (telle que la requête du tableau 1.3 page 10), il est nécessaire de combiner les résultats issus de chacun des Systèmes de Recherche d'Information (SRI) utilisés. Notre problématique principale est de trouver comment réaliser cette combinaison. Néanmoins, comme nous allons le voir par la suite, en RIG, les approches de combinaisons sont peu nombreuses et non flexibles. Un utilisateur ne peut pas paramétrer cette combinaison, par exemple, en favorisant un critère.

Comme nous venons de l'indiquer, nous avons, d'une part, des index contenant des représentations de données et, d'autre part, des méthodes de calcul adaptées à chacune des facettes de l'information géographique. Cette hétérogénéité des représentations et des méthodes de calcul implique la nécessité de les homogénéiser. Il faut donc les uniformiser afin de les combiner comme le préconisent Malczewski et al. [MCF+03] et Pham et al. [PMLC07]. Actuellement, les Systèmes de Recherche d'Information (SRI) classiques

6. Le terme facette désignera l'une des trois composantes géographiques que sont le spatial, le temporel et le thématique.

7. Un critère est une partie de la requête pouvant porter sur une facette géographique. Il faut noter qu'une requête peut contenir plusieurs critères d'une même facette. Par exemple, la requête du tableau 1.3 page 10 contient deux critères spatiaux.

traitent la facette thématique de manière simplifiée par des approches statistiques basées sur les termes. Or, Pham et al. [PMLC07] proposent d'imiter ces approches utilisées pour les termes (troncature, calculs de poids basés sur les fréquences et modèle vectoriel de Salton [Sal71]) afin d'appliquer des adaptations de ces traitements aux images. Nous pensons que les différentes facettes de l'information géographique peuvent aussi être homogénéisées de manière similaire aux approches appliquées aux termes. Par la suite, nous nous limiterons donc à l'étude des approches basées sur le calcul de statistiques.

En recherche d'information classique, les requêtes peuvent contenir plusieurs mots clés. Les moteurs de recherches actuels (tels que Google ou Terrier [OAP+05]) permettent de faire deux types de recherche. La recherche standard se base sur des approches classiques de type TF·IDF et produit scalaire telles que présentés dans le chapitre 2. Ici la requête est constituée uniquement de mots clés. La recherche étendue permet d'ajouter des contraintes sur les différents éléments de la requête. Parmi les opérateurs existants, nous pouvons notamment citer :

- + : exprime une exigence, le terme qui suit l'opérateur doit être présent dans un document résultat ;
- − : exprime une exclusion, le terme qui suit l'opérateur ne doit pas être présent dans un document résultat ;
- ̂ exprime une préférence, cet opérateur associe un coefficient réel qui valorise la présence de ce terme dans un document résultat. Il faut noter que Google n'offre gère pas cet opérateur.

La combinaison des différents éléments de la requête étant facilitée par l'homogénéité de ces derniers (uniquement des mots clés), ces moteurs permettent à un utilisateur de préciser sa requête et de paramétrer de telles combinaisons via des contraintes. Néanmoins, il faut noter que le classement des résultats est souvent opaque. Les moteurs ne spécifient pas, dans l'ensemble résultat présenté, quels sont les critères qui ont été satisfaits et dans quelle mesure ils l'ont été. Nous pensons qu'il est possible d'étendre la combinaison de critères géographiques de manière similaire via des contraintes.

Enfin, concernant la recherche d'information géographique, nous avons pu constater que les systèmes existants n'évaluent que partiellement le gain apporté par la combinaison des différentes facettes de l'information géographique. Notre hypothèse est que la combinaison de ces différentes facettes améliore la pertinence des résultats de manière significative. Néanmoins, comme nous allons le voir par la suite, il n'existe pas de cadre d'évaluation de systèmes de RIG.

1.4 Contributions : uniformisation générique, combinaison personnalisable et évaluations

Dans notre équipe, une chaîne de traitement spatiale permettant de bâtir des index spatiaux et supportant une approche de recherche d'information spatiale a été mise en place dans le prototype PIV par Julien Lesbegueries [Les07]. De la même manière, une chaîne de traitement temporelle générant des index temporels et supportant une approche de recherche d'information temporelle a été mise en place pour le prototype

PIV par Annig Le Parc-Lacayrelle [LGS07]. Pour la facette thématique, il existe de nombreux systèmes de recherche d'information tel que Terrier [8] permettant de travailler sur les termes. Nous disposons donc d'un prototype (PIV) contenant deux chaînes de traitements indépendantes et de SRI dédiés aux termes.

À travers nos différentes contributions nous proposons une alternative aux approches actuellement utilisées en recherche d'information géographique. Ces contributions sont :

1. Une approche d'uniformisation générique que nous appliquons à l'information spatiale ou à l'information temporelle extraite des documents en vue de leur indexation. Il s'agit de mettre en œuvre une stratégie similaire à celles appliquées en RI classique sur les termes (lemmatisation/troncature, calculs de poids basés sur les fréquences et modèle vectoriel de Salton [Sal71]).
2. L'évaluation de la combinaison des différentes facettes de l'information géographique en RI et la quantification de l'apport de cette combinaison. Pour cela, nous proposons, dans un premier temps, d'utiliser des approches linéaires standards ayant fait leurs preuves en RI classique.
3. Une approche de combinaison, originale et générique, basée sur les contraintes et que nous appliquons à la RIG. Le but est de permettre à un utilisateur de personnaliser la combinaison en spécifiant des contraintes pour chaque critère.
4. Un cadre expérimental permettant d'évaluer un SRI géographique.

1.5 Organisation du manuscrit

La partie suivante décrit l'état de l'art sur lequel nous nous sommes appuyés. Dans le premier chapitre de cette partie, sont introduites les notions requises relatives à l'information géographique et la recherche d'information, ainsi qu'un comparatif des systèmes existants. Dans le chapitre suivant, nous présentons différentes méthodes de combinaison existantes, pas nécessairement dédiées à l'information géographique. Pour terminer cette partie, le dernier chapitre illustre différentes approches existantes pour uniformiser des critères avant de mettre en œuvre des stratégies de combinaison.

La troisième partie détaille notre contribution. Dans un premier chapitre, est présentée notre approche générique d'uniformisation, appliquée au spatial ainsi qu'au temporel. Le chapitre qui suit présente nos propositions pour combiner ces différents critères géographiques. Ensuite un chapitre présente les prototypes mis au point, et un dernier détaille nos expérimentations.

La dernière partie contient une synthèse de ce mémoire et propose des perspectives pour la suite de ces travaux.

8. http://ir.dcs.gla.ac.uk/terrier/

Deuxième partie

État de l'Art : de l'information géographique dans des documents textuels à la recherche d'information combinant des critères spatiaux, temporels et thématiques

Introduction de l'état de l'art

Cette deuxième partie s'organise en 3 chapitres. Dans un premier chapitre, nous décrivons les différentes opérations nécessaires à la Recherche d'Information standard mais aussi à la Recherche d'Information Géographique : extraction, indexation, recherche d'information ainsi qu'évaluation. Dans ce premier chapitre les systèmes de RIG les plus représentatifs sont présentés.

Dans un deuxième chapitre, nous nous intéressons à la combinaison de critères. Étant donné que peu d'approches existent en RIG, nous nous sommes intéressés à la combinaison de critères dans d'autres domaines. En Recherche d'Information Multimédia, la fusion de critères permet de combiner des informations provenant de documents de différents types (exemple : images et textes). Pour l'aide à la décision, l'agrégation de critères permet de proposer à un utilisateur les choix les plus proches de ses exigences (tous les critères ne pouvant pas être nécessairement satisfaits en même temps).

Dans un dernier chapitre, nous présentons les différentes approches d'uniformisation existantes et mises en œuvre en amont de la combinaison de critères. La normalisation utilisée en Recherche d'Information permet de borner les scores de pertinences des documents (entre 0 et 1 généralement). La généralisation, utilisée en Recherche d'Information Multimédia, permet de réduire le nombre d'informations en éliminant les détails. La standardisation, pour l'aide à la décision multicritère, permet de convertir des évaluations qualitatives (par exemple : la couleur d'une voiture) en évaluations quantitatives (par exemple : 1 pour bleu et rouge, 0,7 pour orange et jaune, ...). Concernant l'information géographique, il existe une approche de focalisation spatiale qui consiste à réduire l'ensemble des informations spatiales d'un document en une seule.

Chapitre 2

Traitement automatique de l'information géographique dans des documents textuels dans un but de recherche d'information

Sommaire

2.1 Introduction

Dans ce chapitre, nous allons considérer en détail en quoi consiste le traitement de l'information, plus particulièrement de l'information géographique textuelle. Tout d'abord, nous allons définir l'information géographique textuelle. Ensuite les principales modélisations et langages de modélisations géographiques textuels seront exposés. Puis nous expliciterons les différentes étapes liées au traitement automatique de l'information géographique (extraction, indexation, recherche d'information). Pour finir nous évoquerons les principaux Systèmes de Recherche d'Information Géographiques existants. Comme indiqué dans le chapitre précédent, nous nous limiterons à l'étude des approches basées sur les statistiques car nous souhaitons les réutiliser pour les informations spatiales et temporelles.

2.2 L'information géographique dans des textes

Le mot **information** peut avoir diverses significations selon le contexte dans lequel il est utilisé. Au sens étymologique, l'information est l'action de donner une forme. Au niveau du langage, une information est constituée d'une ou plusieurs donnée(s), bien formée(s) et porteuse(s) de sens [Flo09]. La recherche d'information traditionnelle utilise pour représenter l'information contenue dans un document des mots-clés ou plus généralement des termes [9] [BYRN99].

« Selon Goodchild [LGMR05], le problème fondamental de l'information géographique est que celle-ci lie un espace, souvent un instant et quelquefois des propriétés descriptives. Il utilise une métaphore de la chimie en soulignant le caractère atomique des composantes spatiales, temporelles et descriptives de l'information géographique » [Lou08]. L'**information géographique**, peut donc se définir comme un ensemble de trois facettes : thème, espace et temps [Use96, Gai01, Lou08]. Elle peut se représenter sous différentes formes : représentation graphique (pour le spatial par exemple en 2D (carte) ou 3D (avec les élévations)), représentation textuelle (sous forme d'expression) ou encore représentation sous forme de données (tuples dans une base de données).

Dans notre cas, nous travaillons sur l'information géographique représentée sous forme textuelle. Cette information est donc diluée dans le discours, ce qui rend difficile son extraction. Par exemple, dans l'extrait suivant : « Le 8 août 1903, Henri Russell accomplit sa trente-troisième ascension et redescend avec tristesse à Gavarnie. » (tableau 1.1 page 9), un lieu est mentionné (Gavarnie) mais il n'est pas précisé s'il s'agit de la commune, du Cirque ou encore de la station de ski. La figure 2.1 illustre l'information géographique avec un exemple textuel. Dans cet exemple, l'information spatiale est

9. Un terme est un mot ou groupe de mots ayant du sens. Il est qualifié de mot-clé lorsqu'il a été présélectionné [BYRN99].

représentée par « au sud de Pau ». Ce syntagme [10] permet de retrouver uniquement les documents traitant du « sud » et de « Pau ». Cette même information spatiale représentée par une géométrie 2D pourra retourner beaucoup d'informations (documents évoquant Jurançon, Gan, ...) grâce à des opérateurs spatiaux adaptés (tels que la translation et l'intersection). Cette limite est aussi vraie pour les autres facettes. Pour le temporel, « XX^{e} siècle » représentée par un intervalle de temps permet de retourner toutes les dates ou périodes qui s'y rapportent (par exemple : 1905, été 1960, ...).

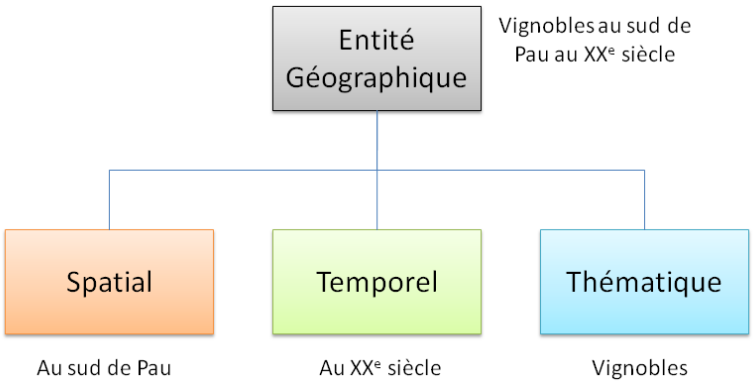


FIGURE 2.1 – Information Géographique

Ces trois facettes doivent toujours exister. Néanmoins dans une unité documentaire (dans notre cas le paragraphe), certaines facettes peuvent ne pas être présentes ou de manière implicite. Par exemple, une information temporelle peut être indiquée dans un paragraphe et ne pas être répétée dans ceux qui suivent ou uniquement de manière partielle.

Pour ne pas se restreindre à ces représentations textuelles, il est donc nécessaire d'identifier ces informations géographiques et de les convertir en données permettant de tirer parti de leur spécificité. Un traitement basé sur une analyse sémantique du texte permet de détecter les informations spatiales (ou temporelles) d'un document et de leur associer une représentation symbolique (tel que « au sud de Pau » est une représentation de type orientation appliquée à la commune de Pau). Néanmoins, pour pouvoir réaliser, lors de la recherche, des opérations spatiales (calcul d'intersection par exemple), il est nécessaire de calculer une représentation numérique. Les informations détectées peuvent être subjectives ou dépendantes du contexte d'invocation, donc les représentations numériques associées impliquent toujours une certaine approximation.

Ainsi, de manière générale, l'information spatiale détectée dans un syntagme nominal est successivement représentée sous forme textuelle, symbolique et enfin numérique (voir figure 2.2). La validation et l'approximation numérique d'une telle information spatiale

10. Un syntagme est un regroupement de mots. C'est donc une unité intermédiaire entre le mot et la phrase [RPR99]

nécessite l'usage de bases de connaissances particulières : dictionnaires spatiaux (gazetteers) pouvant être manipulés via des outils dédiés tel que les Systèmes d'Information Géographiques (SIG). Un gazetteer est une liste de noms de lieux associés à leur localisation (coordonnées). A ces lieux peuvent être aussi précisées diverses caractéristiques (par exemple statistiques tels que la population, ou physiques tels que le relief). Prenons l'exemple du gazetteer Geonames [11], chaque entrée est décrite par un nom, un pays, un type (parc, lac, montagne, ville, ...), une latitude et une longitude. Le tableau 2.1 montre les propriétés de la ville de Pau sur Geonames. Un système d'information géographique permet d'une part de stocker des données spatiales, et, d'autre part d'utiliser des opérateurs pour les manipuler (intersection, distance, ...). Les données spatiales peuvent être plus ou moins précises : uniquement des points (latitude/longitude par exemple), seulement les coordonnées du rectangle délimitant l'information spatiale (on parle de boîte englobante ou MBR pour *Minimum Bounding Rectangle* en anglais), ou encore la forme géométrique fine (tel qu'un polygone) (voir tableau 2.2).

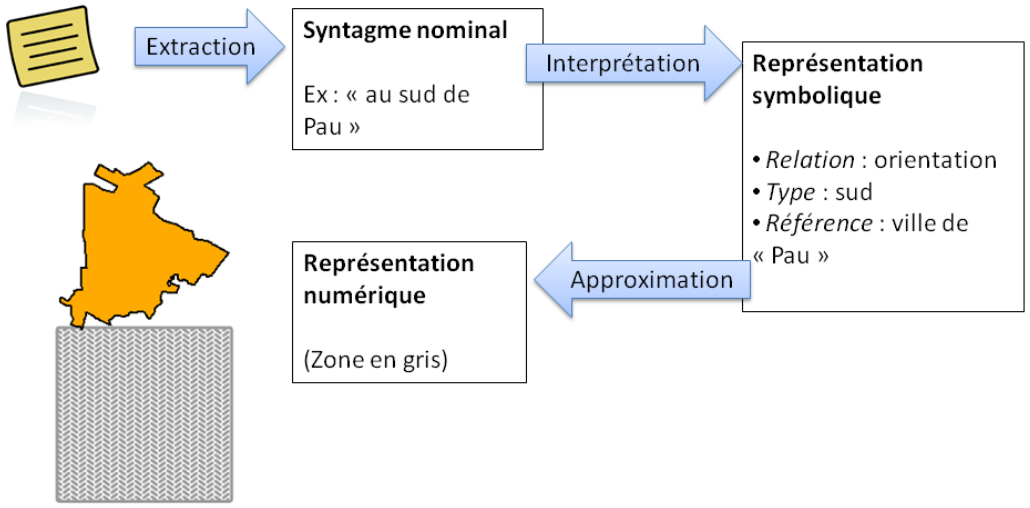


FIGURE 2.2 – Traitement de l'information spatiale

Propriété	Valeur
Nom	Pau
Pays	France, Aquitaine
Classe	lieu habité, population 82 697
Latitude	N 43°18' 0"
Longitude	W 0°22' 0"

TABLE 2.1 – Entrée « Pau » dans Geonames

11. http://www.geonames.org/

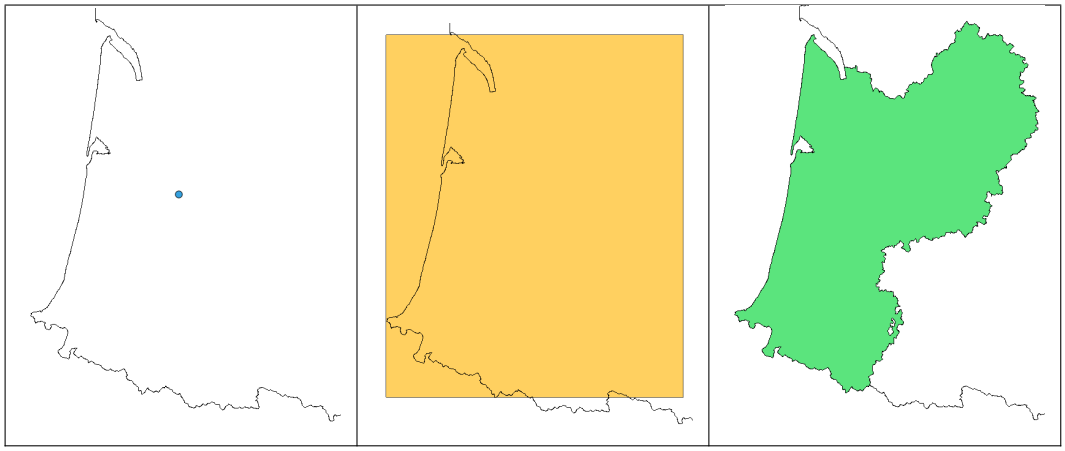


TABLE 2.2 – Exemple de représentations possibles pour Aquitaine (respectivement : centroïde, boîte englobante (MBR) et polygone)

Pour l'information temporelle, le principe est le même : détectée dans un syntagme nominal, elle est successivement représentée sous forme textuelle, symbolique puis numérique (ici ce sont des intervalles de temps et non des points ou géométries). Pour traiter l'information calendaire il est aussi nécessaire de disposer de bases de connaissance, néanmoins moins complexes que pour le spatial.

Enfin concernant la facette thématique, l'information reste généralement limitée aux termes utilisés en recherche d'information standard. Néanmoins des termes différents peuvent couvrir des thèmes identiques (exemple : automobile et voiture). Cette approche peut être complétée par des ressources externes (thésaurus, ontologies) contenant des liens de synonymie ou hiérarchiques (voir figure 2.3). Nous envisageons la combinaison des facettes spatiales, temporelles et thématiques. Toutefois pour le thématique, nous utiliserons les modèles et outils de RI classiques. Aussi, ne nous détaillerons pas davantage la facette thématique qui se limitera à l'exploitation des termes.

Maintenant que nous avons présenté l'information géographique dans des documents textuels, nous allons nous intéresser aux travaux relatifs à la modélisation de cette information.

2.3 Modélisation de l'information géographique exprimée dans des documents textuels

L'information géographique, de par sa spécificité, nécessite l'usage d'une modélisation adaptée à chacune de ses facettes. Les traitements appliqués pour extraire l'information géographique de discours textuels étant limités, les modèles utilisés sont généralement succincts et formels. Dans ce contexte, pour le spatial, un modèle de référence est RCC-8 (*Region Connection Calculus*) [RCC92,Les07] qui définit huit relations entre deux régions x et y, telles que le recouvrement partiel ou l'égalité (voir figure 2.4 pour les différentes

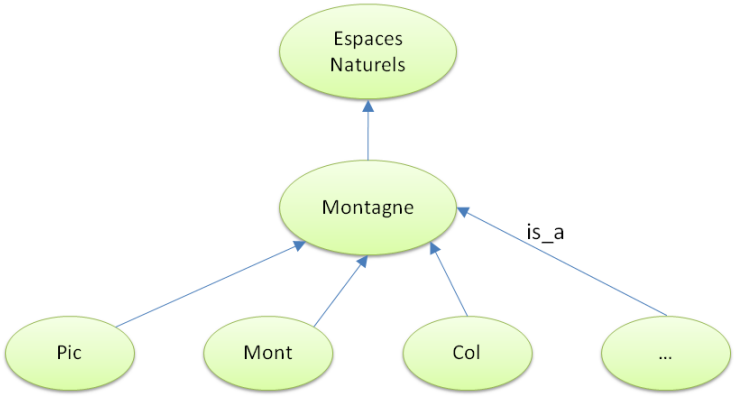


FIGURE 2.3 – Exemple de liens hiérarchiques pouvant être exprimés dans une ontologie

relations topologiques). Il existe des extensions permettant de prendre en compte les représentations linéaires [EMH94]. Pour le temporel, un modèle de référence est celui proposée par Allen mettant en œuvre les relations entre intervalles de temps [All84, MT04] (voir figure 2.5 pour les différentes relations temporelles).

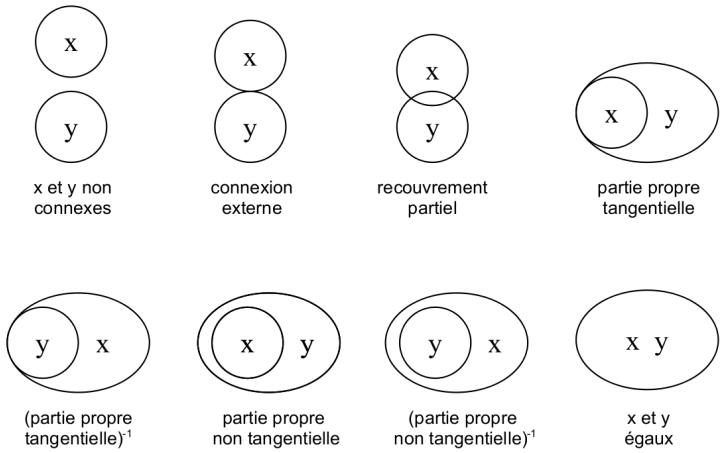


FIGURE 2.4 – Les 8 relations topologiques pouvant exister entre 2 régions x et y selon le modèle RCC-8 [RCC92] (illustration extraite de [Les07])

Concernant les langages de modélisation pour l'information géographique textuelle, nous pouvons distinguer plusieurs types en fonction de leur finalité : échange ou description des connaissances. La plupart sont réalisées en XML (*eXtensible Markup Language*), qui est un langage de balisage générique permettant de structurer l'information [BB99]. Dans cette section, nous prendrons l'exemple de l'information spatiale « au sud de Pau » pour illustrer les différents marquages spatiaux. De même, pour le temporel nous utili-

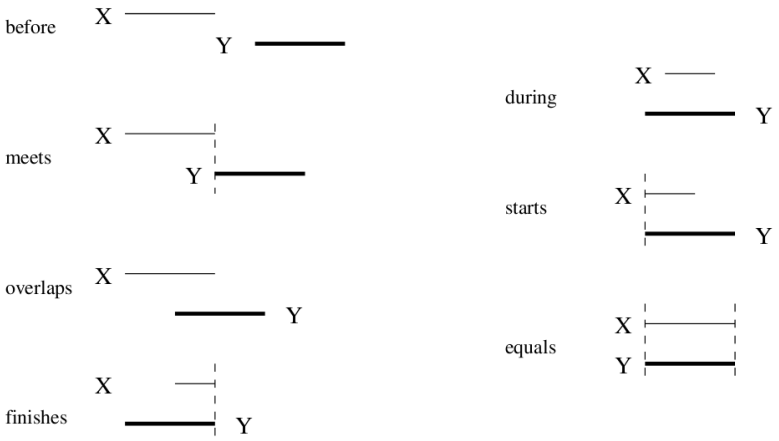


FIGURE 2.5 – Relations d'Allen [All84] (illustration extraite de [MT04])

serons l'exemple « début de janvier 2010 ». Ces représentations numériques nécessitent d'être calculées.

2.3.1 Langages de modélisation pour l'information spatiale

Un langage de modélisation spatial de type format d'échange très répandu est le *Geography Markup Language* (GML). Définit par l'OGC [12], il permet de stocker des objets géographiques, plus particulièrement les géométries correspondantes (représentations numériques). GML a par ailleurs été conçu pour être utilisé dans d'autres langages XML. Il gère uniquement les représentations numériques (donc pas de représentations symboliques). Le listing 2.1 illustre le code GML de la représentation « au Sud de Pau ». Comme nous pouvons le voir, nous avons un objet de type polygone (« <gml:Polygon> ») et les latitudes/longitudes de ses différents points.

Listing 2.1 – Exemple de GML (au sud de Pau)

```
<gml:Polygon>
<gml:outerBoundaryIs>
<gml:LinearRing>
<gml:coordinates>
 -0.389339593262433,43.2345070972552
 -0.392743259810513,43.3061317796098
 -0.294231809315081,43.3085863498989
 -0.290950779544868,43.2369584558224
 -0.389339593262433,43.2345070972552
</gml:coordinates>
</gml:LinearRing>
</gml:outerBoundaryIs>
</gml:Polygon>
```

12. L'*Open Geospatial Consortium* (OGC) est un consortium international proposant des formats ouverts sur l'information géographique

Keyhole Markup Language (KML) [13], est un autre langage de modélisation spatial de type format d'échange très répandu. Défini également par l'OGC, il est notamment utilisé dans *GoogleMaps* ou *GoogleEarth*. Tout comme le GML, il gère uniquement les représentations numériques, par contre il peut décrire des styles associés aux représentations (couleur, épaisseur des bordures, ...). Le listing 2.2 illustre le code KML de la représentation « au Sud de Pau ». Comme dans l'exmple du GML listing 2.1, nous avons un objet de type polygone (« <Polygon> ») et les latitudes/longitudes de ses différents points. La principale différence est qu'aux coordonnées sont associées des informations pour la visualisation (couleur rouge, trait épais).

Listing 2.2 – Exemple de KML (au sud de Pau)

```
<kml>
<Document>
<Style id="redLine">
<LineStyle><color>ff0000ff</color><width>4</width></LineStyle>
</Style>
<Placemark>
<styleUrl>#redLine</styleUrl>
<Polygon>
<outerBoundaryIs>
<LinearRing>
<coordinates>
-0.389339593262433,43.2345070972552
-0.392743259810513,43.3061317796098
-0.294231809315081,43.3085863498989
-0.290950779544868,43.2369584558224
-0.389339593262433,43.2345070972552
</coordinates>
</LinearRing>
</outerBoundaryIs>
</Polygon>
</Placemark>
</Document>
</kml>
```

SpatialML [14] [MHR+08] est un langage de marquage spatial de type description des connaissances. Il a été développé par l'organisation américaine MITRE [15]. SpatialML gère les représentations numériques des lieux (balise PLACE). Par contre, pour les relations spatiales (balises SIGNAL et LINK), il ne stocke que des représentations symboliques (voir figure 2.2 page 22 pour les différents types de représentations). Le listing 2.3 illustre le code SpatialML de la représentation « au Sud de Pau ». Comme nous pouvons le voir, la ville de Pau a été identifiée et des coordonnées lui ont été associées ; la relation d'orientation (sud) a aussi été identifiée mais il n'y a pas de représentation numérique associée.

Listing 2.3 – Exemple de SpatialML (au sud de Pau)

```
<SIGNAL id="1" type="DIRECTION">sud</SIGNAL>
<PLACE id="2" country="FR" form="NAM" latlong="43.301667N -0.368611W">Pau</
    PLACE>
```

13. http://www.opengeospatial.org/standards/kml/
14. https://spatialml.mitre.org/
15. http://www.mitre.org/

```
<PLACE id="3" />
<RLINK id="4" distance=2 direction="S" source="2" target="3" signals="1"/>
```

2.3.2 Langages de modélisation pour l'information temporelle

Pour l'information temporelle, le langage de modélisation textuel le plus répandu est TIMEX3 (successeur de TIMEX2). Il permet de représenter numériquement des informations temporelles au format standard ISO-8601 [Man03]. Les listing 2.4 illustre le code TIMEX3 de la représentation « début de janvier 2010 ».

Listing 2.4 – Exemple de TIMEX3 (début de janvier 2010)

```
<TIMEX3 tid="t2" type="DATE" value="2010-01-10" />
<TIMEX3 tid="t3" type="DURATION" value="P15D" beginPoint="t1" endPoint="t2"
    />
```

Pour le marquage temporel, il existe un équivalent à SpatialML : TimeML [16] [PCI+03, PKLS05]. Il utilise TIMEX3 pour le marquage des données temporelles. Tout comme SpatialML il marque les relations temporelles mais ne leur associe que des représentations symboliques (pas de représentations numériques). Le listing 2.5 illustre le code TimeML de la représentation « début de janvier 2010 ». Comme nous pouvons le voir, la date « janvier 2010 » a été identifiée et une représentation numérique lui a été associée ; la relation temporelle d'inclusion (début) a aussi été identifiée mais il n'y a pas de représentation numérique associée.

Listing 2.5 – Exemple de TimeML (début de janvier 2010)

```
<SIGNAL sid="s1">debut</SIGNAL>
de
<TIMEX3 tid="t1" type="DATE" value="2010-01">
janvier 2010
</TIMEX3>
<TLINK eventInstanceID="ei1" relatedToTime="t1" signalID="s1" reltype="
    BEGINS"/>
```

Il existe des langages dédiés aux représentations numériques des informations que ce soit pour le spatial (GML) ou le temporel (TIMEX3). Les langages SpatialML et TimeML ont été mis en place pour pouvoir intégrer ces représentations numériques et y ajouter les relations spatiales et temporelles. Ces informations étant floues, il est intéressant de conserver les représentations symboliques. Néanmoins, ces langages ne permettent pas d'associer des représentations numériques à ces représentations symboliques. À noter que contrairement à SpatialML qui est mis à jour régulièrement (version 3 diffusée en octobre 2009), TimeML semble être arrêté actuellement (version 1.2.1 sortie en 2005).

Afin de pouvoir marquer l'information comme nous venons de le voir, il faut d'abord l'identifier via une phase d'extraction. Puis une fois passées ces deux étapes, ces informations pourront être indexées afin de permettre à un utilisateur de faire des recherches.

16. http://www.timeml.org/site/index.html

2.4 Extraction et Indexation dans un but de Recherche d'Information Géographique

Dans cette partie, nous allons expliquer les différentes étapes nécessaires à la Recherche d'Information. Étant donné que nous nous intéressons surtout à la Recherche d'Information Géographique, la Recherche d'Information traditionnelle ne sera présentée que de manière succincte.

Il existe un grand nombre de systèmes supportant la recherche d'information tels que : Lemur [17] [OC01], Lucene [18] [GH05], ou Terrier [19] [OAP+05]. Les moteurs de recherche, tel que Google ou Yahoo utilisent aussi des systèmes de recherche d'information similaires. Pour la recherche d'information sémantique, il existe des plateformes de traitement automatique du langage naturel (TALN) tel que GATE [20] [CMBT02] ou LinguaStream [21] [BW06]. Pour l'information géographique, il existe des projets qui se sont intéressés à une ou plusieurs facettes. Nous allons à la fin de ce chapitre (section 2.5) voir en détail les plus représentatifs.

2.4.1 Extraction d'Information dans un but de Recherche d'Information Géographique

Dans [Poi03], Poibeau a mis en place un glossaire dans lequel il donne cette définition de l'extraction d'information : « Activité qui consiste à remplir automatiquement une banque de données à partir de textes écrits en langue naturelle [Paz97, Paz99]. Elle s'oppose classiquement à la recherche documentaire (information retrieval), qui vise à retrouver dans une base de documents un ensemble de documents pertinents au regard d'une question [SM83, Voo99] ». Or ici nous allons aborder l'extraction d'information dans le but de constituer des index à des fins de recherche d'information. En effet, pour accélérer la recherche et donner des scores de pertinence aux documents il est indispensable de passer par une phase d'extraction et d'indexation. Dans un premier temps, nous allons présenter de l'extraction d'information standard, puis nous traiterons les cas particuliers du spatial et du temporel.

2.4.1.1 Extraction d'information classique

L'analyse lexicale d'un document permet de convertir un flux de caractères en flux de mots ou termes [BYRN99]. Dans cet ensemble de termes, ne sont conservés que ceux qui sont significatifs. Certains termes peu discriminants (ex : « de », « à ») sont éliminés grâce à ce qu'on appelle une liste de mots vides (ou stoplist). Une fois l'ensemble des termes discriminants extraits, un processus dénommé lemmatisation [GGHR00], permet d'obtenir pour chaque terme son lemme. Un lemme est la forme canonique d'un mot.

17. http://www.lemurproject.org/
18. http://lucene.apache.org
19. http://ir.dcs.gla.ac.uk/terrier/
20. http://gate.ac.uk/
21. http://www.linguastream.org/

Par exemple, pour un verbe c'est son infinitif. Ainsi pour chaque document, la phase d'extraction d'information permet d'obtenir l'ensemble des lemmes qu'il contient. Le tableau 2.3 présente un extrait des lemmes obtenus pour le texte du tableau 1.1. Pour pouvoir lemmatiser les termes et éliminer les termes peu discriminants, un système doit donc utiliser un algorithme adapté à la langue et une liste de mots vide pour chaque langue. Certains systèmes ne supportent que quelques langues (pour Lemur : anglais, chinois et arabe). Néanmoins, la lemmatisation n'est pas la seule méthode existante. Par exemple, la troncature raccourcit les mots dépassant une certaine taille fixée (par exemple : avantages → avantag pour une taille fixée à sept caractères). Les termes extraits seront ensuite pondérés selon des formules que nous présentons dans la section 2.4.2.1.

Token	Conservé	Lemme
Pendant	oui	pendant
que	non	
Russel	oui	russel
courait	oui	courir
le	non	
monde	oui	monde
...		

TABLE 2.3 – Exemple de résultat de lemmatisation du texte du tableau 1.1 avec le logiciel TreeTagger [Sch94].

Des traitements avancés cherchent à extraire les entités nommées présentes dans des documents textuels. Les **entités nommées** représentent « l'ensemble des noms de personnes, d'entreprises et de lieux présents dans un texte donné. On associe souvent à ces éléments d'autres syntagmes repérables par des grammaires locales comme les dates, les unités monétaires, les pourcentages... » [Poi03]. L'extraction d'entités nommées nécessite au préalable une phase d'extraction telle que présentée ci-dessus pour les termes. Ensuite, un détecteur d'entités nommées (NER pour *Named Entity Recognition* en anglais) utilise généralement une base de règles ou un système d'apprentissage pour détecter les entités nommées et les catégoriser [Poi03]. Voici un exemple de règle : « un nom propre précédé par la préposition à, est potentiellement un lieu ». Le système GATE cité auparavant contient un module de détection d'entités nommées : Annie [22]. Le système *LinguaStream* [23] permet également de construire et d'appliquer de telles règles.

2.4.1.2 Extraction d'information spatiale

L'extraction d'information spatiale nécessite d'utiliser un détecteur d'entités nommées spatial tels que MetaCarta [24] ou OpenCalais [25]. Ici, seuls les lieux nous intéressent. Ce traitement permet d'obtenir une liste d'entités spatiales candidates. Ces entités spa-

22. http://gate.ac.uk/ie/annie.html
23. http://www.linguastream.org/
24. http://www.metacarta.com
25. http://www.opencalais.com

tiales identifiées peuvent être comparées à une ou plusieurs bases de données spatiales afin d'être validées. L'interrogation de ces bases de données spatiales permet aussi de récupérer toutes les informations disponibles sur ces entités spatiales (types, coordonnées, ...). Une des plus répandue et facile d'accès est Geonames [26]. Néanmoins elle ne contient que les coordonnées ponctuelles des lieux. Il existe d'autres bases de données plus précises telles que celles fournies par l'IGN (Institut Géographique National) [27], contenant les polygones précis des communes ou départements ou encore tous les pics, refuges et autres points d'intérêts sous forme de ponctuels (BD TOPO ®, BD NYME ®). Ainsi les informations spatiales contenues dans ces bases sont considérées comme valides, et, il est possible de récupérer des données sur ces informations : latitude/longitude ou encore le polygone.

Afin de calculer les représentations des relations spatiales, il est nécessaire de réaliser des opérations sur les coordonnées obtenues (via des opérateurs, tel que la translation). En effet, pour l'entité « près de Pau », il n'y a pas de représentation spatiale bien définie. Il est néanmoins possible de calculer et de proposer des approximations [SGPL08]. Pour cette entité, une fois que « Pau » a été détecté via un traitement de type NER, l'expression « près de Pau » peut être reconstruite via des outils de traitement automatique de la langue naturelle (TALN).

Cette phase de validation des informations spatiales pose néanmoins des problèmes : de nombreux lieux portent le même nom ou des noms de lieux sont utilisés pour représenter des concepts tels que des institutions (par exemple « Quai d'Orsay » pour le ministère des Affaires Etrangères). Dans [MMB08], Martins et al. citent une étude [EI05] montrant que 67 % des toponymes sont ambigus dans un document. Plus le gazetteer couvre une surface vaste et détaillée, plus le nombre d'ambigüités est élevé [LSSS07]. Il existe néanmoins diverses solutions, par exemple en étudiant la proximité des entités spatiales d'un document, pour lever les ambigüités [LSSS07] ou encore via des méthodes probabilistes prenant en compte des paramètres tels que l'importance du lieu, sa population, ... [LMSC06].

2.4.1.3 Extraction d'information temporelle

Les documents textuels contiennent plusieurs types d'informations temporelles : l'information calendaire (sous forme de date plus ou moins complète) et les entités nommées (tels que « Renaissance » ou « 2^e guerre mondiale ») [MMBS09, BL04] par exemple.

L'information temporelle est généralement extraite avec un détecteur d'entités nommées comme pour le spatial. Néanmoins, les expressions temporelles sont souvent incomplètes et diffuses dans un document [TM08]. En effet, un auteur peut énoncer une date complète, puis par la suite omettre l'année car le lecteur sait que cela se déroule durant la même année. De plus, selon le type de corpus l'ordonnancement temporel peut être différent comme le montrent Mani et al. [MSZ03] avec les actualités qui ont un ordonnancement inversé (actualité la plus récente en premier). Certains travaux ont donc

26. http://www.geonames.org/
27. http://www.ign.fr/

cherché à déterminer les liens entre les différentes informations temporelles consécutives afin de les ordonnancer ainsi que préciser celles qui sont incomplètes [MSZ03, MT04].

Concernant les entités nommées temporelles, il est nécessaire de posséder une base de données contenant ces références et les dates associées. Néanmoins, tout comme le spatial, le temporel est concerné par le problème d'ambigüité [BL04] : si nous prenons l'entité « Renaissance », elle ne correspond pas à la même période selon les pays (en France de 1494 à 1610, en Angleterre de 1520 à 1620, ...). Des projets de gazetteer spatiotemporel [MMBS09, BGL07] se mettent en place afin de répondre à ce type de problème.

Nous avons donc vu dans cette partie l'extraction d'information dans un but d'indexation pour la recherche d'information standard ainsi que pour le spatial et le temporel. Une fois les informations extraites des documents, elles vont donc être stockées dans des index afin de supporter la recherche d'information.

2.4.2 Indexation d'Information dans un but de Recherche d'Information Géographique

À moins que les documents soient vraiment très courts ou changent constamment, il n'est pas envisageable de les parcourir dans leur intégralité à chaque recherche [BYRN99]. L'indexation consiste donc à créer une structure de données, nommée index, ne contenant que les informations importantes d'un document en vue de le retrouver. Ainsi l'indexation permet de conserver le résultat d'une phase d'extraction et de le réorganiser afin d'économiser des ressources et du temps dans des phases de Recherche d'Information ultérieures.

Concernant l'information géographique qui est une combinaison de 3 facettes (espace, temps et thème), la question de la séparation ou non des 3 index est importante. De nombreux travaux [MSA05, VJJS05, LSSS07] suggèrent que chaque facette doit être indexée indépendamment des autres. Cela permet de traiter une facette sans tenir compte des autres, de simplifier la gestion et la modification de chaque index, et de rendre la combinaison éventuelle de facettes en phase de RI plus flexible.

Nous allons maintenant présenter l'indexation en RI, puis de l'indexation spatiale et enfin temporelle.

2.4.2.1 Indexation standard

La phase d'extraction produit une liste de lemmes par document. Suite à ce traitement, un index de lemmes va alors être produit. Pour chacun de ces lemmes indexés, un poids est affecté afin de supporter des classements de résultats par la suite. Comme expliqué dans [Bes04], le poids peut être une combinaison de plusieurs pondérations : locales, globales et normalisations. La pondération locale la plus utilisée est le TF (*Term Frequency*) qui est la fréquence d'apparition du lemme dans le document [SYY75]. Si le lemme est cité 5 fois dans le document, sa fréquence est de 5. Il existe d'autres pondérations locales telles que : le facteur binaire (1 si le terme est présent dans le document 0

sinon), le facteur logarithmique (basé sur le TF, il cherche à compenser le trop fort poids des termes trop fréquents), le facteur augmenté (borne le TF des termes présents entre 0,5 et 1). Concernant la pondération globale, elle est utilisée pour mesurer l'importance d'un lemme dans la base documentaire : la plus utilisée est l'IDF (*Inverse Document Frequency*) [SYY75]. Cela permet de minorer le poids d'un lemme présent dans toute la base par exemple, car il sera peu discriminant. Pour pallier le fait que ces mesures ne tiennent pas compte de certains critères, tel que la longueur du document pouvant influer sur la fréquence des termes, des normalisations ont été mises en place (normalisation par le cosinus, max-TF, normalisation à pivot et la normalisation unique à pivot) [Bes04].

Le système Lucene par exemple permet de paramétrer la pondération. Il propose par défaut une formule dérivée du TF·IDF [SD08], mais aussi d'utiliser le TF et/ou IDF classiques, de les normaliser, ou encore de rajouter des formules. Il est aussi possible de favoriser les termes contenus dans un titre en augmentant leurs poids.

La plupart des systèmes mettent en place des index classant les lemmes par ordre alphabétique qui se nomment Index Inversé. Le tableau 2.4 présente un exemple d'index inversé pour lequel la pondération utilisée est le TF seul.

Lemmes	Documents	TF
pendant	para441	1
Vignemale	para446 para461	1 3
...		

TABLE 2.4 – Exemple d'index inversé

2.4.2.2 Indexation spatiale

Un index spatial peut être similaire à un index standard illustré précédemment : un index inversé classant les termes spatiaux auxquels sont associés les documents dans lesquels ils apparaissent comme dans [LMSC06, MSA05]. Il peut aussi stocker des informations propres au spatial comme les coordonnées (latitude/longitude) ou les géométries [VJJS05, LGL06]. Les systèmes de gestions de bases données spatiales contiennent des structures particulières pour gérer ces coordonnées ou géométries afin d'accélérer les recherches : R-Tree, B-Tree ou encore GiST. L'intérêt d'intégrer les représentations des informations spatiales dans l'index est de pouvoir, par la suite, calculer spatialement leurs similarités avec les informations spatiales exprimées dans des requêtes d'utilisateurs (calcul d'intersection, de distance, ...).

2.4.2.3 Indexation temporelle

Tout comme le spatial, l'index temporel peut s'apparenter à un index de termes, c'est à dire une liste inversée d'informations temporelles liées à des documents, ou, à un index plus riche contenant des informations temporelles telles que des représentations (intervalles de temps) [LGS07].

Comme nous avons pu le voir, selon le type d'information, un index peut contenir des représentations numériques directement exploitables par des opérateurs de RI. Une fois ce ou ces index générés, ils vont pouvoir être interrogés pour sélectionner les documents correspondants aux requêtes des utilisateurs.

2.4.3 Recherche d'Information Géographique (RIG) dans les documents textuels

Gérard Salton, l'un des pionniers de la recherche d'information, la définit comme un domaine s'intéressant à la structuration, l'analyse, l'organisation, le stockage, la recherche et la récupération de l'information [Sal68, CMS09]. Comme l'illustre la figure 2.6, un système de recherche d'information doit gérer trois processus (boites arrondies) : la représentation du contenu des documents (*indexing*), la représentation du besoin d'information de l'utilisateur (*query formulation*) et la comparaison entre ces deux représentations (*matching*) [GD09]. Si les résultats ne correspondent pas à son besoin, l'utilisateur va pouvoir modifier sa requête (*feedback*).

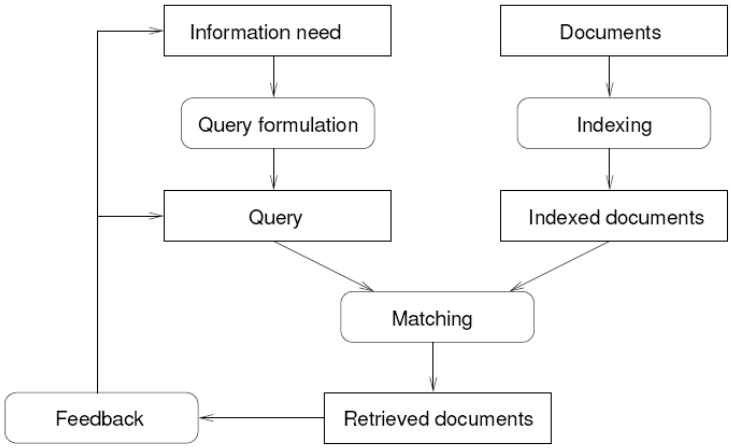


FIGURE 2.6 – Processus de recherche d'information (illustration extraite de [GD09])

La recherche d'information géographique peut être assimilée à une branche spécialisée de la recherche d'information traditionnelle [Lar96].

Tous les traitements présentés auparavant sont réalisés en amont, une seule fois dans le cas de bibliothèques numériques car les documents n'évoluent pas. Pour des sites web cela est différent, étant donné qu'ils évoluent en continu, il faut actualiser l'index régulièrement. La recherche d'information va donc consister à interroger le ou les index générés afin de retourner des résultats classés par rapport à une requête soumise par un utilisateur. Généralement, la requête subit le même traitement qu'un document, à la différence qu'à la fin les informations extraites ne sont pas stockées mais sont comparées aux index pour déterminer les documents pertinents.

Une requête, en recherche d'information, est l'expression d'un besoin d'information d'un utilisateur dans le langage proposé par le système d'information. Le langage de requêtage par mots-clés est le plus répandu [BYRN99]. En plus des mots-clés, il est possible d'utiliser des opérateurs booléens (ET, OU, SAUF). Par exemple sur Google, par défaut les mots sont séparés par des ET, pour l'opérateur de disjonction OU il faut mettre « OR » et pour le sauf il faut utiliser « − ». Nous allons, ici aussi, aborder en premier la RI en général, puis la RI spatiale et temporelle.

2.4.3.1 Recherche d'information standard

En recherche d'information, l'utilisateur soumet une requête composée de mots clés, ces derniers sont lemmatisés et comparés aux index [BYRN99]. Plusieurs modèles existent pour réaliser cette comparaison et mesurer la similarité entre les informations de la requête et de chaque document de l'index. Parmi les plus répandus figurent le modèle booléen et le modèle vectoriel [Flu04]. Le modèle booléen [LF73], permet d'avoir un ensemble de documents pertinents et un ensemble de documents non pertinents, seul le premier étant retourné aux utilisateurs. Néanmoins, les résultats ne sont pas ordonnés. C'est la raison pour laquelle le modèle vectoriel [Sal71] est très utilisé [Bes04]. Il consiste à représenter l'ensemble des termes d'un document sous la forme d'un premier vecteur ainsi que l'ensemble des termes de la requête sous la forme d'un deuxième vecteur et de comparer ces 2 vecteurs. Cela revient à représenter la base documentaire sous la forme d'une matrice, comme sur la figure 2.5 (D correspond à document, T à terme, et w_{ij} au poids du terme j pour le document i).

$$\begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ D_n \end{array} \begin{array}{c} \begin{array}{cccc} T_1 & T_2 & \dots & T_t \end{array} \\ \begin{pmatrix} w_{11} & w_{21} & \dots & w_{t1} \\ w_{21} & w_{22} & \dots & w_{t2} \\ \vdots & \vdots & & \vdots \\ w_{n1} & w_{n2} & \dots & w_{tn} \end{pmatrix} \end{array}$$

TABLE 2.5 – Modèle vectoriel : matrice document-par-termes

Pour mesurer la similarité entre deux vecteurs il existe plusieurs mesures [Bes04, CMS09] : de type vectoriel tels que le produit scalaire (produit des deux vecteurs) ou le cosinus de l'angle entre les deux vecteurs (intersection normalisée), de type ensembliste tels que la mesure de similarité de Dice (rapport de l'intersection sur la moyenne arithmétique des normes) ou la mesure de similarité de Jaccard (rapport de l'intersection sur l'union) ; la mesure la plus utilisée étant la mesure du cosinus (voir équation 2.1). Ce modèle permet de récupérer en résultat une liste de documents pertinents classés par

ordre de pertinence. Ce modèle est très répandu. Il a néanmoins quelques limites : il considère que chaque terme est indépendant, ce qui n'est pas le cas en réalité [Bes04].

$$\text{sim(A,B)} = cos(\theta) = \frac{A \cdot B}{||A||\,||B||} \tag{2.1}$$

2.4.3.2 Recherche d'information spatiale

Contrairement à la recherche d'information classique où une requête utilisateur prend la forme d'un ou plusieurs mots clés, elle peut avoir différentes formes pour le spatial. Il est possible de fournir les coordonnées d'un point (latitude/longitude) [LSSS07], de dessiner la zone souhaitée sur une carte [VJJS05, Val06] ou, comme cité auparavant, de l'exprimer sous forme de mots clés ou phrases en langue naturelle [LGL06]. Une requête en texte libre subit un traitement préalable afin d'en extraire l'information spatiale (comme cela est effectué pour l'extraction et l'indexation des informations spatiales dans les documents).

Une fois l'information spatiale extraite de la requête, il faut déterminer quels documents contiennent des informations pertinentes pour cette requête et leur donner un score. Le calcul de pertinence spatiale, c'est à dire de similarité spatiale entre la requête et une information d'un document se fait en règle générale par un calcul d'intersection ou de distance entre les représentations numériques de la requête et du document [JP08]. Nous pouvons notamment citer les mesures de similarités suivantes [LF04, And10] :

- Distance euclidienne : classe les documents selon la proximité entre les représentations spatiales de la requête et du document.
- Degré de recouvrement : classe les documents selon la surface de recouvrement entre les représentations spatiales de la requête et du document ; plus elle est grande, plus le document est pertinent.
- Relations de confinements : classe les documents selon le ratio d'inclusion entre les représentations spatiales de la requête et du document.

Le score d'un document est généralement calculé avec des méthodes linéaires [MSA05] telles que la moyenne arithmétique [BDEH07] ou le maximum [SBLG07].

2.4.3.3 Recherche d'information temporelle

Tout comme pour le spatial, une requête temporelle peut s'exprimer sous différentes formes. En règle générale, elle est exprimée sous forme textuelle [BDEH07, LGS07]. Certains travaux s'intéressent à des formes d'interrogations graphiques : Googlelabs expérimente un prototype avec une ligne de temps comme indiqué par [NRD08]. Cela permet de ne pas avoir à extraire l'information temporelle de la requête.

Une fois l'information temporelle extraite de la requête, il faut déterminer quels documents contiennent des informations pertinentes pour cette requête et leur donner un score. Le calcul de pertinence temporelle, c'est à dire de similarité temporelle entre la requête et une information d'un document se fait en règle générale par un calcul d'intersection ou de distance entre les représentations numériques de la requête et du

document [BDEH07,LGS07]. Comme pour le spatial, le score d'un document est généralement calculé avec des méthodes linéaires telles que la moyenne arithmétique [BDEH07] ou le maximum [LGS07].

Nous avons vu ici les approches utilisées pour la Recherche d'Information standard et spécialisée (spatiale et temporelle). Pour terminer cette section, nous allons nous intéresser à l'évaluation des systèmes de RI.

2.4.4 Évaluation d'un Système de Recherche d'Information Géographique

Comme expliqué dans [San10], il est important d'évaluer un système de RI, pour informer l'utilisateur sur la qualité du système et pouvoir faire évoluer ce système sur des critères objectifs.

Il existe de nombreuses mesures pour évaluer un système de RI [Flu04]. Les plus utilisées sont la précision et le rappel (*precision* et *recall* en anglais). La précision est le nombre de documents pertinents parmi les documents retournés. Le rappel est le nombre de documents pertinents retournés par le système parmi l'ensemble des documents pertinents. Cela permet de mesurer à quel point le système retourne des documents pertinents. D'autres mesures basées sur ces deux là existent et sont très utilisées telle que la mesure de précision moyenne. La précision moyenne (*Average Precision*) calcule la précision d'une requête donnée sur l'ensemble de ses résultats. La MAP (MAP pour *Mean Average Precision* en anglais) calcule la moyenne des précisions moyennes des différentes requêtes afin d'avoir une vue d'ensemble des résultats du système. Le tableau 2.6 illustre les formules des mesures qui viennent d'être présentées.

Le domaine de la RI est caractérisé par une longue tradition d'évaluation, notamment au travers du cadre TREC (*Text REtrieval Conference*) [VH05] qui permet l'évaluation des SRI au regard de la facette thématique. De par la maturité de ce domaine, les SRI donnent de bonnes performances (*efficiency*) au niveau du stockage ou du temps de calcul par exemple. Ainsi, de nous jours, les campagnes comme TREC portent sur la qualité des résultats (*effectiveness*). Par ailleurs, la facette temporelle a également fait l'objet du cadre d'évaluation TempEval [VGS+09]. Bucher et al [BCJ+05] ont proposé de considérer deux facettes simultanément : spatiale et thématique. Cette proposition se retrouve dans les tâches GeoCLEF [GLS+05] et GikiCLEF [SC10] du cadre CLEF (*Cross Language Evolution Forum*) [Pet01]. Ce dernier a notamment permis, à partir de requêtes géographiques, l'évaluation de SRI thématiques classiques en RI tels que Lemur [OC01], Lucene [GH05] et Terrier [OAP+05], comme rapporté dans [POGCGVUL08]. Quand à GeoTime [GLK+10], il propose d'utiliser les facettes spatiales et temporelles. A notre connaissance il n'existe donc pas de cadre d'évaluation traitant toutes les facettes de l'information géographique. La RIG étant un domaine plus récent, les SRIG font face à des problèmes de performance (*efficiency*), c'est la raison pour laquelle les campagnes portant sur l'information géographique ne couvrent pas toutes ses facettes.

Pour calculer ces mesures, il faut au préalable avoir une base de référence, dans laquelle des experts ont identifié les documents pertinents pour une requête. Ce travail est

Précision	$P(s,r) = \frac{Perti(r) \cap Restit(s,r)}{Restit(s,r)}$
Rappel	$R(s,r) = \frac{Restit(s,r)}{Perti(r)}$
Average Precision	$AP(s,r) = \frac{\sum_{rank=1}^{Restit(s,r)} P(s,r,rank) * rel(rank)}{Perti(r)}$
Mean Average Precision	$MAP(s) = \frac{\sum_{r=1}^{Nb_r} AP(s,r)}{Nb_r}$
$Perti(r)$: Nombre de documents pertinents pour la requête r, $Restit(s,r)$: Nombre de documents restitués par le système s pour la requête r, Nb_r : Nombre de requêtes effectuées $rel(rank)$: fonction binaire qui retourne 1 si le résultat est pertinent	

TABLE 2.6 – Formules utilisées pour évaluer un système de RI

donc fastidieux, d'autant que pour réaliser une évaluation correcte, elle doit avoir lieu sur un corpus suffisamment représentatif (plusieurs milliers de documents). Aujourd'hui les campagnes d'évaluation utilisent des corpus trop importants pour être évalués manuellement (TREC1 contient 1Go de données). Comme l'illustre la figure 2.7 [Voo07], TREC recourt à la technique du *pooling* [SJvR75]. Ainsi, pour chaque requête (*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.

Concernant les SRIG, les travaux existants (qui seront présentés dans la section suivante) ont tout au plus été évalués du point de vue de la taille des index générés et du temps de réponse. 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 facettes de l'information géographique de ce point de vue. Il est donc impossible de comparer les moteurs de recherche qui sont capables de gérer les trois facettes de l'information géographique.

Évaluer un système de RI est donc une tâche importante et nécessaire pour déterminer la qualité de ses résultats. Néanmoins, l'évaluation manuelle est toujours nécessaire à titre de comparaison, ce qui restreint les évaluations d'un système pour lequel les cam-

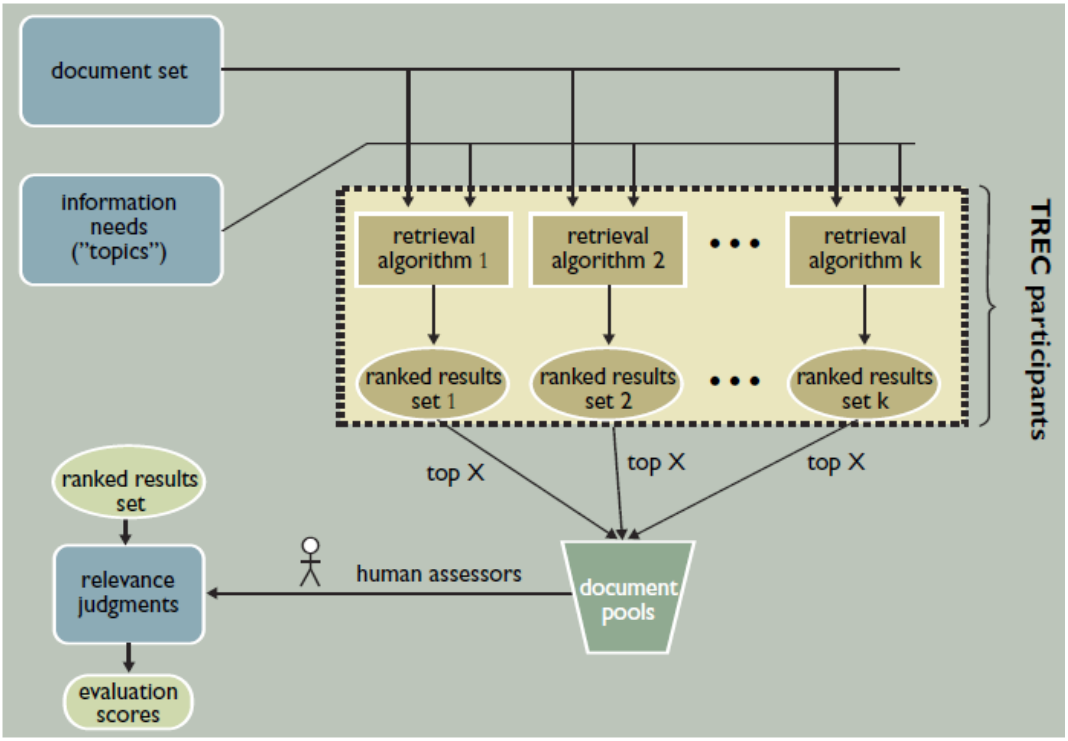


FIGURE 2.7 – Evaluation d'un SRI (illustration extraite de [Voo07])

pagnes d'évaluations actuelles ne seraient pas adaptées. Maintenant nous allons détailler plusieurs systèmes dédiés à la RIG.

2.5 Systèmes dédiés à la Recherche d'Information Géographique

Le tableau 2.7 présente les principaux Systèmes de Recherche d'Information Géographique existants selon les critères suivants :

- **Type de documents.** La place prépondérante d'Internet aujourd'hui, ainsi que sa taille en perpétuelle expansion, font que la plupart des systèmes lui sont dédiés (STEWARD [LSSS07], SPIRIT [VJJS05] mais aussi GRID [Val06] ou GéoTracker [CDG+07] qui a la particularité de s'intéresser aux flux RSS). Néanmoins, certains systèmes visent des types de documents différents : GéoSem traite des documents plus ou moins longs (plusieurs pages) [BDEH07], DIGMAP [MMBS09, MBP+07] gère des cartes anciennes, CITER [PEH+09] et PIV [LGL06] travaillent sur des livres anciens (plusieurs centaines de pages).
- **Documents restitués.** La plupart des systèmes renvoient le même type de document qu'ils acceptent en entrée. Par contre, traiter des documents longs apporte

	PIV [LGL06]	**STEWARD** [LSSS07]	**SPIRIT** [VJJS05]	**GéoSem** [BDEH07]	**DIGMAP** [MMBS09]	**CITER** [PEH+09]
Types de documents	Livres	Textes courts	Pages web	Textes longs	Cartes historiques	Livres
Documents restitués	Paragraphes de livres	//	//	Extrait de texte	//	Sections
Spatial	lieux + relations	lieux	lieux + (relations)	lieux + relations	lieux	lieux
Temporel	dates + relations	-	-	dates + relations	dates	dates
Thème	Termes	Termes	Termes	Termes	Termes	Thème
Représentations spatiales	Polygones	MBR	MBR	MBR	Points	Points
Interrogation	Langage naturel	Mots clés + lieu (optionnel)	Mots clés + lieu + relation spatiale (3 champs)	3 champs : spatial, temporel et termes	chaque facette a 1 (ou +) champs	3 champs : spatial, temporel et thème

Table 2.7 – Systèmes de Recherche d'Information Géographique

des contraintes supplémentaires : en effet, il n'est pas envisageable de restituer un livre complet lorsqu'un utilisateur recherche une information précise. C'est la raison pour laquelle, les systèmes manipulant ce genre de documents proposent des points d'entrées particuliers tel que le paragraphe pour PIV [LGL06], une section de chapitre pour CITER [PEH+09] ou des extraits pour GéoSem.

– **Spatial.** La majorité des systèmes traitent la facette spatiale. Néanmoins, la plupart se limitent aux informations toponymiques (lieux). Le système SPIRIT [VJJS05] travaille sur des adresses postales (des lieux précis) et gère les relations spatiales exprimées dans des champs spécifiques de l'interface d'interrogation (il ne les gère pas au niveau de l'indexation). Le système STEWARD [LSSS07], utilise deux gazetteers (GNIS – Geographic Names Information System pour les USA et GNS - GEOnet Names Serveur pour le reste du monde) mais n'a pas accès à des opérateurs spatiaux pour traiter des relations spatiales. Le système GéoSem [BDEH07], malgré l'absence de SIG, gère certaines relations spatiales (orientation, proximité) définies par des règles. L'utilisation d'un SIG dans la plateforme PIV [LGL06], a permis d'envisager de nombreuses relations spatiales pour approximer les informations floues et d'en construire une représentation géoréférencée.
– **Temporel.** Peu de systèmes traitent du temporel. Les systèmes DIGMAP [MMBS09] et CITER [PEH+09] gèrent uniquement les dates, tandis que les systèmes GéoSem [BDEH07] et PIV [LGS07] traitent à la fois les dates et les relations temporelles. Dans ces systèmes, les informations temporelles sont représentées par des

intervalles de temps (une date de départ, une date de fin). Les systèmes ne traitent que des informations calendaires (dates, périodes).

- **Thème.** Seul CITER [PEH+09] utilise des concepts provenant d'une ontologie. L'utilisateur doit sélectionner un thème dans la liste disponible mais ne peut pas fournir d'autres mots-clés comme c'est le cas dans les autres systèmes. Ces derniers utilisent des approches statistiques standards de RI sur les termes.
- **Représentation spatiale.** Concernant les représentations des informations spatiales, la plupart des systèmes travaillent avec des boîtes englobantes (MBR). Seuls DIGMAP [MMBS09] et CITER [PEH+09] se basent sur des données ponctuelles, et PIV [LGL06, SGPL08] sur des données géométriques (polygones).
- **Interrogation.** Afin de simplifier le traitement des différentes facettes de la requête, certains systèmes demandent aux utilisateurs de les exprimer séparément dans une interface d'interrogation adaptée. C'est le cas de GéoSem [BDEH07], qui propose trois champs à remplir (spatial, temporel, mots clés), idem pour CITER [PEH+09], SPIRIT [VJJS05] et STEWARD [LSSS07] (spatial et termes uniquement pour ces deux derniers). SPIRIT propose en plus de choisir une relation spatiale à appliquer sur l'information spatiale demandée, tel que « au sud de ». GéoSem gère aussi les relations spatiales qu'il est capable de détecter dans la requête spatiale, idem pour le temporel (trois champs implique trois requêtes). Concernant le prototype PIV [LGL06], il accepte des requêtes multifacettes qui seront traitées par les différentes chaînes qui ne feront ressortir que leur facette respectives (par exemple, la chaîne spatiale n'extraira que les informations spatiales).

Nous avons constaté que beaucoup de systèmes ne traitent que deux facettes, généralement le spatial et le thématique (limité aux termes) : dans le tableau 2.7 c'est le cas de SPIRIT [VJJS05] et STEWARD [LSSS07], mais parmi les travaux existants nous pouvons citer aussi GRID [Val06], GéoSVM [Cai02] et Zettair [LMSC06]. Quelques systèmes travaillent sur les trois facettes : GéoSem [BDEH07], DIGMAP [MMBS09], WatWasWaar [LG08], CITER [PEH+09], PIV [LGL06]. Nous allons aborder la combinaison des résultats de RI dans le chapitre suivant. Néanmoins, si de nombreux travaux préconisent la combinaison ou proposent des approches possibles [Cai02, MSA05], peu l'ont mis en place. De plus, cela se limite généralement à une intersection sans score global.

2.6 Conclusion

Ce chapitre a permis d'introduire les notions liées à la recherche d'information, que ce soit de manière classique telle que dans les moteurs de recherches usuels, ou pour l'information géographique. Traiter de l'information géographique est plus compliqué que des termes car il faut considérer plusieurs facettes (spatial, temporel et thématique). Étant donné que le contenu des pages Web change continuellement, ces techniques peuvent s'avérer trop coûteuses. Néanmoins, leur contenu étant généralement court et grâce aux progrès matériels et logiciels réguliers, ces traitements pourront aussi être plus largement utilisés. Dans notre cas, nous travaillons sur des fonds stables (le contenu d'un livre ne

change pas au cours du temps), utiliser ce genre de traitements plus lourds est moins problématique.

Le traitement distinct et indépendant des différentes facettes de l'information géographique est préconisé [CJP06, MSA05]. Néanmoins, cela implique qu'il faut ensuite les combiner durant la phase d'interrogation, ce qui pose un certain nombre de problèmes. En effet, pour chacune des facettes, les index sont précis mais très spécifiques et les représentations stockées très hétérogènes. Les résultats des techniques de RI dédiées à ces facettes ne sont donc pas homogènes en terme de mode de calcul de scores de pertinence notamment. Nous allons ensuite nous intéresser aux approches et opérateurs de combinaisons de résultats de RI. En Recherche d'Information Géographique ces approches sont encore peu nombreuses, nous nous sommes alors intéressés à ce qui se pratique dans d'autres domaines.

Chapitre 3

Combinaison de critères

Sommaire

3.1 Introduction

Parmi les systèmes de Recherche d'Information Géographique existants, peu traitent les 3 facettes et les combinent. Aussi, nous sommes nous également intéressés aux approches de combinaisons utilisées dans des domaines tels que la Recherche d'Information Multimédia et l'aide à la décision. Il faut néanmoins préciser que, selon le domaine, le terme employé n'est pas « combinaison » car même si la finalité est d'unir plusieurs critères, le contexte, la quantité de critères et les méthodes s'avèrent très différents. Pour la RI multimédia, il s'agit de « fusion de critères » et pour l'aide à la décision, « d'agrégation de critères ». La fusion de critères permet de rassembler les ensembles d'informations extraites de différents documents (par exemple : images et textes) ou de différentes parties d'un même document (par exemple une vidéo peut se décomposer en un ensemble d'images et une bande sonore). L'agrégation de critères est utilisée pour aider un utilisateur à faire son choix en trouvant les meilleurs compromis (souvent tous les critères ne peuvent pas être satisfaits). En Recherche d'Information Géographique, les approches de combinaison sont généralement de type filtrage, consistant à interroger un SRI sur chacune des facettes successivement puis à réaliser l'intersection des résultats.

Dans ce chapitre, le terme « combinaison » englobera les différentes approches présentées ici (fusion, agrégation, filtrages ou de type linéaires). Nous allons donc voir en quoi consistent ces différentes approches ainsi que celles utilisées en Recherche d'Information classique.

3.2 Fusion et Recherche d'Information Multimédia

La mise en place de services comme Youtube [28] ou Dailymotion [29], permettant à tous les utilisateurs de mettre en ligne facilement et gratuitement des vidéos et de les partager avec le reste du monde, fait que le nombre de documents multimédias croît de manière exponentielle (Youtube a annoncé que 24 heures de vidéos étaient mises en ligne toutes les minutes [30]).

De nombreux travaux s'intéressent à l'extraction directe d'information depuis ces documents. Une vidéo peut être décomposée en un ensemble d'images, auquel s'ajoute une piste audio comme l'illustre la figure 3.1. Il est ainsi possible d'extraire des informations de chaque image (exemple : couleurs, textures, ...), de la piste audio (par exemple le discours), ou encore en comparant deux images successives (par exemple par rapport aux mouvements). À moins que l'utilisateur fasse une recherche via des images (par exemple il donne une image d'une voiture pour en trouver d'autres de la même voiture), ces informations sont trop peu porteuses de sens pour des recherches classiques par mot-clés. Il est donc nécessaire de déterminer quels concepts sont contenus dans une image ou une vidéo [LSDJ06]. Un concept est une information porteuse de sens, par exemple un chat, un sport, un bus, le ciel, ... La « généralisation » permet la conversion des informations « brutes » extraites d'un document en concepts. Sur la figure 3.1, la généralisation est représentée par un entonnoir (en haut à droite) prenant en entrée différents types de données (couleurs, textures, ...) et renvoyant des concepts. Sur cette figure, nous pouvons voir que des concepts sont issus de la piste audio (chat, sport) et des images (bus, ciel). Cette généralisation sera abordée plus en détail dans le chapitre suivant.

La fusion consiste à rassembler des ensembles d'information de même nature, tels que des concepts. Pour une vidéo cela consiste à fusionner ceux extraits des images et ceux extraits de la piste audio [AGQ06]. Sur notre exemple (figure 3.1), l'ensemble obtenu contient donc : bus, ciel, chat et sport. Cette approche peut aussi fusionner des informations extraites d'une image et du texte qui lui est associé, par exemple dans une page Web [PMLC07].

Cette approche de fusion permet donc de générer de nouveaux index de concepts. Les informations directement extraites des documents (couleurs, textures, ...) peuvent être conservées dans des index dédiés pour permettre d'autres approches de recherche (recherche par image par exemple).

Pour réaliser une telle fusion, il est nécessaire que les informations extraites de chaque document ou partie de document passent par une phase de généralisation afin d'obtenir un ensemble de concepts. Il faut noter que cette fusion se fait de manière équitable : si une image et un texte sont fusionnés, aucun des deux n'est avantagé. En effet dans les différents travaux cités aucun ne pondère les différents types de données utilisés.

28. http://www.youtube.com/
29. http://www.dailymotion.com
30. http://youtube-global.blogspot.com/2010/03/oops-pow-surprise24-hours-of-video-all.html

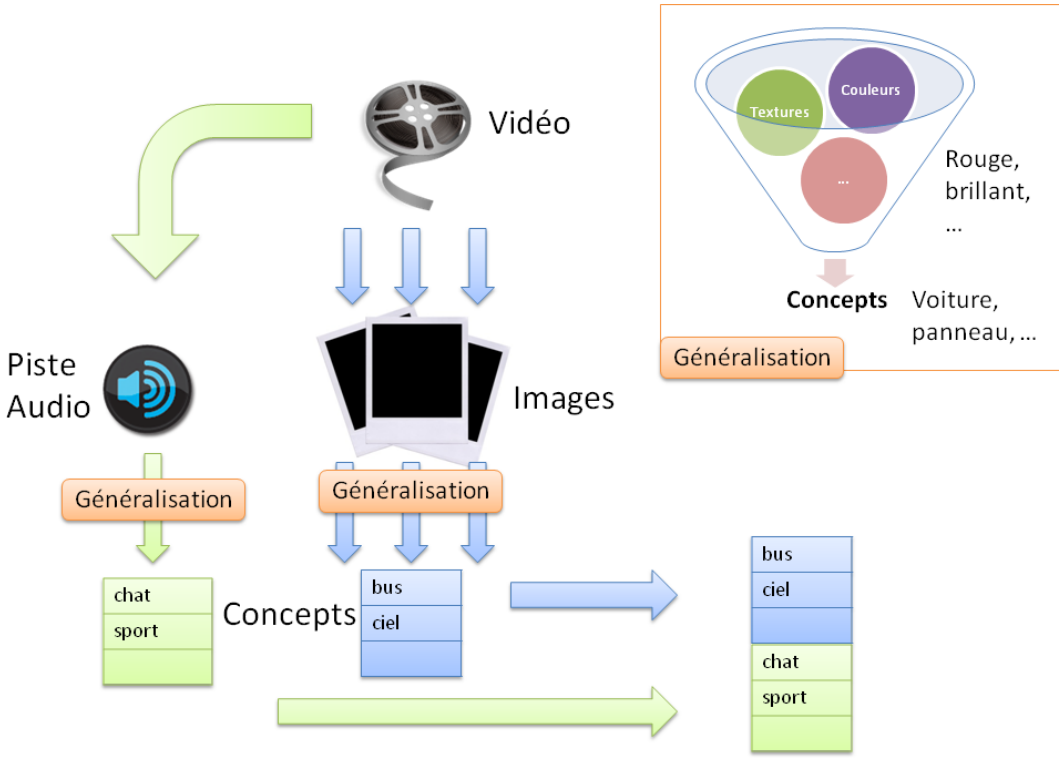


FIGURE 3.1 – Fusion sur une vidéo

3.3 Agrégation de critères et Systèmes d'aide à la Décision

Les Systèmes d'Aides à la Décision (SAD ou DSS pour Decision Support Systems) sont utilisés depuis une quarantaine d'années pour faciliter la prise de décision [Pow97]. Très vite ces systèmes se sont orientés vers une approche multicritère [BPP+93] (plusieurs critères à satisfaire) car la « réalité » est multidimensionnelle [Zel82]. L'objectif d'une aide multicritère à la décision est de proposer à un utilisateur les meilleures solutions à son problème en sachant que toutes les contraintes ne pourront pas nécessairement être respectées et qu'il faudra donc faire des choix.

Un problème de décision multicritère peut être formulé par le modèle « A, A/F, E » [Mar99]. A est l'ensemble des alternatives, c'est à dire des choix possibles. A/F est l'ensemble des attributs ou critères (*Attribute/Feature*), c'est à dire des contraintes exprimées par l'utilisateur. E est l'ensemble des évaluations de performances des alternatives selon chacun des critères, c'est à dire à quel point l'alternative répond au critère. La figure 3.2 illustre un problème d'aide à la décision : la construction d'un nouveau lotissement. L'utilisateur spécifie ses critères (A/F), puis, le système classe les différentes alternatives (A) s'approchant le plus de ses besoins par rapport aux différentes évaluations (E) effectuées au préalable (les v et x du tableau de droite).

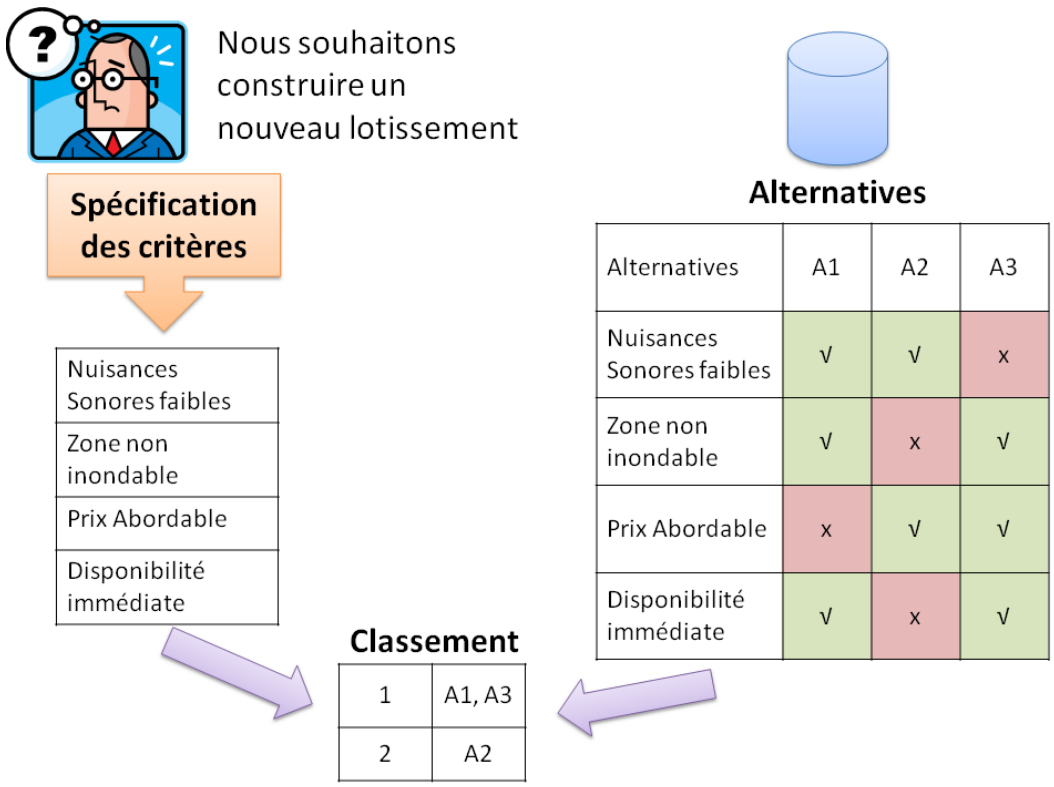


FIGURE 3.2 – Agrégation de critères

Comme l'illustre la figure 3.2, deux alternatives peuvent avoir la même position (A1 et A3) alors que des critères différents sont satisfaits. Pour favoriser certains critères, il est possible de spécifier des préférences [MCF+03, BR10]. La figure 3.3 ajoute les préférences à la figure 3.2 ce qui influence le classement des alternatives. Sur ces deux figures, les alternatives sont évaluées de manière binaire (satisfait ou non satisfait). Or en règle générale, l'évaluation est réalisée de manière proportionnelle (score compris entre 0 et 1).

La figure 3.4 illustre le même exemple avec des évaluations quantitatives proportionnelles (sans utiliser les préférences). Ici nous constatons que bien que A2 satisfasse uniquement deux critères mais de manière forte (0,9) il est classé au même niveau que A3 qui satisfait trois critères mais de manière moins importante. Comme le soulignent certains travaux [Yag88, MCF+03], le modèle initial (sans préférences) ne permet pas de moduler la compensation des critères. Un tel système où tous les critères sont équivalents permet à une alternative satisfaisant moins de critères qu'une autre (tel que A3 et A2) d'être classée aussi bien voire mieux grâce à un meilleur score sur les critères satisfaits. Les préférences permettent donc de favoriser certains critères mais aussi d'influencer leur compensation. La figure 3.5 illustre toujours le même exemple, où l'utilisateur a mis un

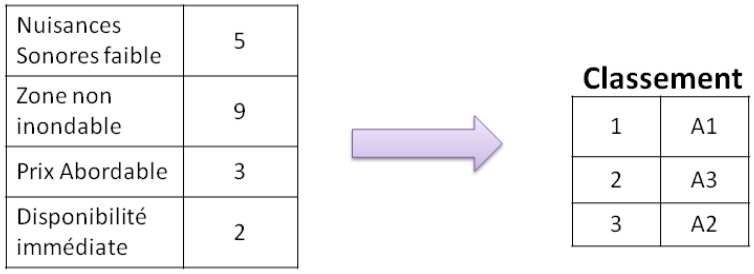


FIGURE 3.3 – Agrégation de critères (avec préférences)

poids important à « Zone inondable » et un poids faible à « Disponibilité » et « Prix » pour éviter que ces derniers ne compensent le premier. Par contre le modèle ne prévoit pas d'exiger la présence d'un critère.

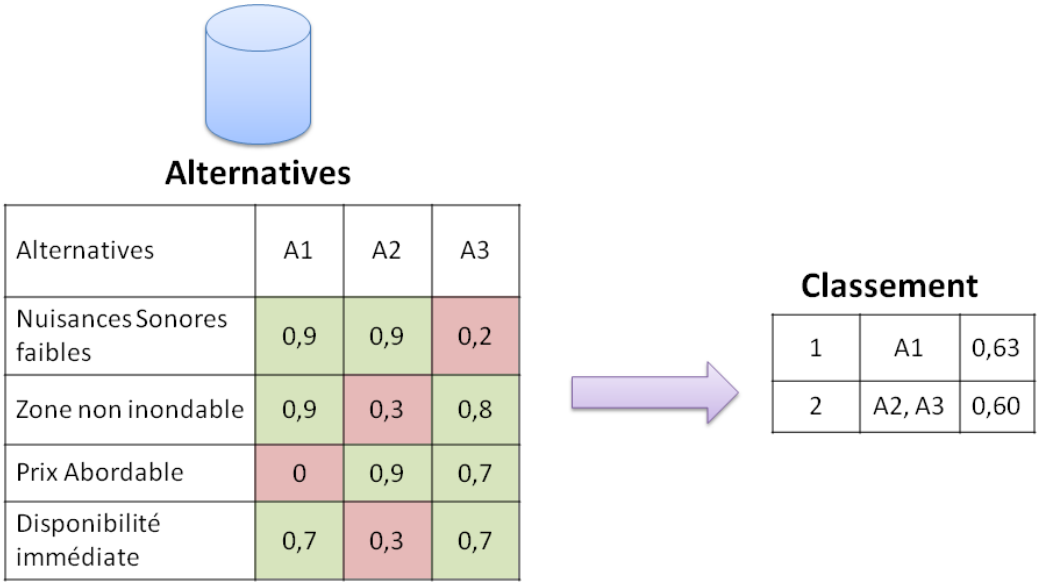
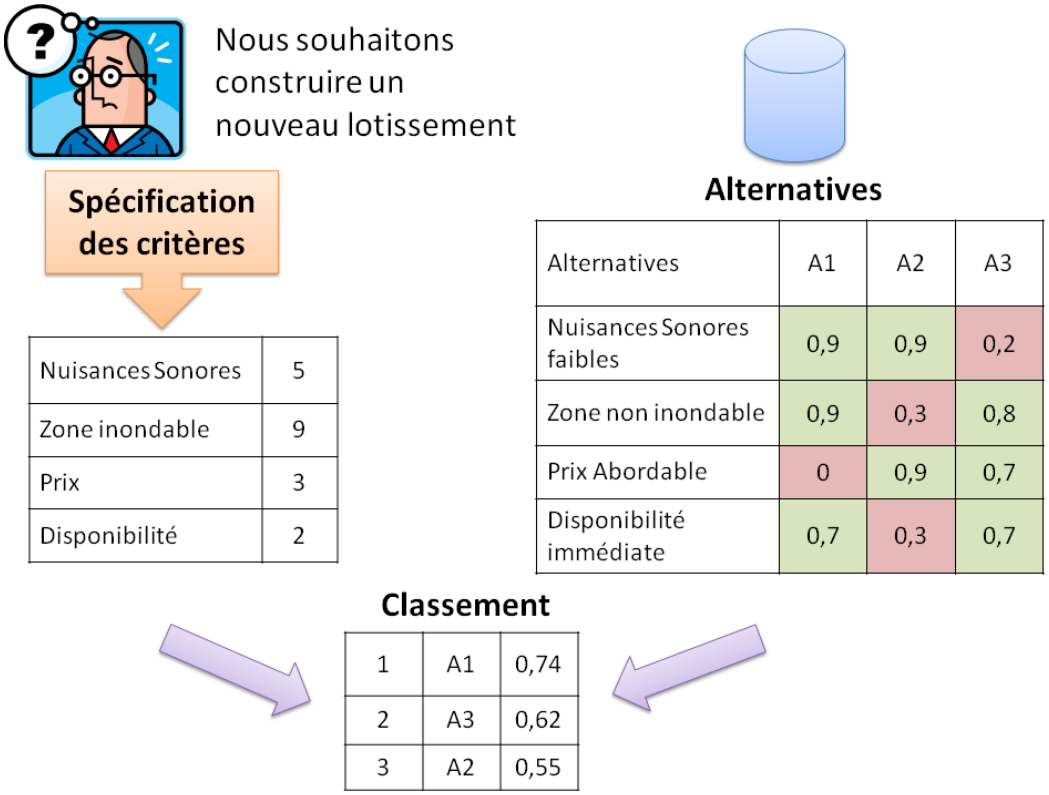


FIGURE 3.4 – Agrégation de Critères (avec évaluations quantitatives proportionnelles)

Il existe des opérateurs permettant à un utilisateur non pas de spécifier directement les préférences, mais le degré de compensation entre les critères et ainsi générant les préférences adéquates tel que la méthode de la moyenne pondérée (OWA pour *Ordered Weighted Averaging*) [Yag88]. Deux cas extrêmes sont distingués : tous les critères doivent être satisfaits (AND), un seul critère satisfait suffit (OR). L'objectif du nouvel opérateur OWA est de relier ces deux extrêmes. Les opérateurs de type « *anding* » sont appelés t-norms (par exemple la fonction Min). Les opérateurs de type « *oring* » sont appelés co-



Nuisances Sonores	5
Zone inondable	9
Prix	3
Disponibilité	2

Alternatives	A1	A2	A3
Nuisances Sonores faibles	0,9	0,9	0,2
Zone non inondable	0,9	0,3	0,8
Prix Abordable	0	0,9	0,7
Disponibilité immédiate	0,7	0,3	0,7

1	A1	0,74
2	A3	0,62
3	A2	0,55

FIGURE 3.5 – Agrégation de Critères (avec évaluations quantitatives proportionnelles et préférences)

t-norms (par exemple la fonction Max). L'utilisation d'un « *anding* » permet de ne faire aucune compensation entre critères, alors qu'un « *oring* » fait une compensation totale entre les attributs en ne prenant en compte que le meilleur attribut. Or, en règle générale un utilisateur cherche un intermédiaire entre ces deux principes. L'opérateur OWA ou « *orand* » permet d'ajuster le degré de « *anding* » et de « *oring* ». Cela consiste à trier les scores des critères par ordre decroissant (par exemple : si on a les critères $A = 0,2$, $B = 0,7$ et $C = 0,5$ alors le vecteur est [B,C,A]) et à définir des poids pour chaque rang et non pas par critère (ex : le premier élément du vecteur de critères aura un poids de 0,4, le 2^e 0,3), le total des poids ne devant pas dépasser 1. Cet opérateur respecte les propriétés d'un opérateur d'agrégation : la monotonicité et la symétrie (grâce à l'ordonnancement inverse des scores des critères). Ce vecteur est borné inférieurement par « *oring* » (de type [1;0;0], tel que Max), et supérieurement par « *anding* » (de type [0;0;1], tel que Min). Il est donc possible de se rapprocher de l'un ou de l'autre en jouant sur les poids. Pour obtenir la moyenne il suffit d'utiliser comme poids 1/NombreElements (par ex : [1/3;1/3;1/3]). Ces poids associés aux rangs peuvent être calculés automatiquement en choisissant l'écart souhaité avec l'une des bornes [MCF+03]. La figure 3.6 illustre les principaux cas de

l'opérateur OWA (compensation totale, compensation moyenne, aucune compensation). Comme nous pouvons le voir, le réglage de la compensation modifie complètement les résultats : par exemple A1 qui est en tête dans les 2 cas de compensations, n'apparaît pas dans les résultats s'il n'y a aucune compensation (car le critère « Prix » n'est pas du tout satisfait).

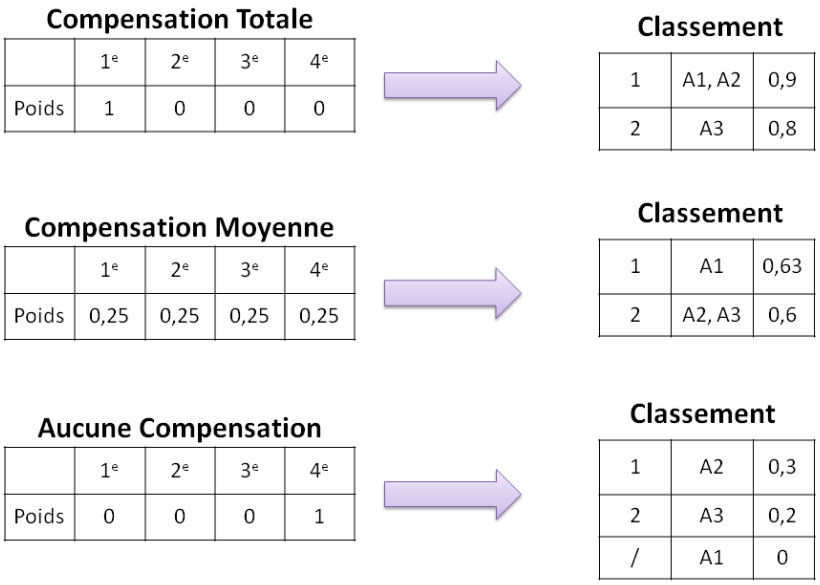


FIGURE 3.6 – Agrégation de Critères (avec l'opérateur OWA)

D'autres travaux cherchent à atténuer la compensation, comme la méthode OWA, mais aussi à améliorer cette agrégation via un système de priorité. Costa Pereira et al. [CPDP09a] proposent d'utiliser quatre critères pour évaluer la pertinence d'un document : sujet (*aboutness*), couverture, justesse, fiabilité. En plus d'un poids associé à chacun de ces critères, ils leurs associent un ordre de priorité. Par exemple, le sujet est le critère avec la priorité numéro 1 : s'il n'est pas satisfait, les autres critères ne sont pas examinés. Le premier critère doit donc être obligatoirement satisfait. La figure 3.7 illustre deux cas selon l'ordre des critères choisi. Pour le deuxième cas, l'alternative A1 est éliminée car elle ne satisfait pas le critère obligatoire (le premier).

Toutes les approches présentées considèrent que les différents critères sont indépendants. La moyenne arithmétique pondérée utilisée dans la plupart de ces approches ne permet pas de prendre en compte les relations existantes entre critères. Afin d'intégrer les liens existants entre certains critères (par exemple l'implication, de type $crit_A \rightarrow crit_B$), plusieurs travaux proposent d'utiliser des méthodes adaptées telle que l'intégrale de Choquet [BR10,LG03,HL09]. Par exemple, Büyüközkan et al. [BR10] expliquent que dans le cadre de l'évaluation des risques en développement de logiciels, certains risques (critères) sont liés. Au lieu d'associer un poids à un critère, l'intégrale de Choquet donne des poids

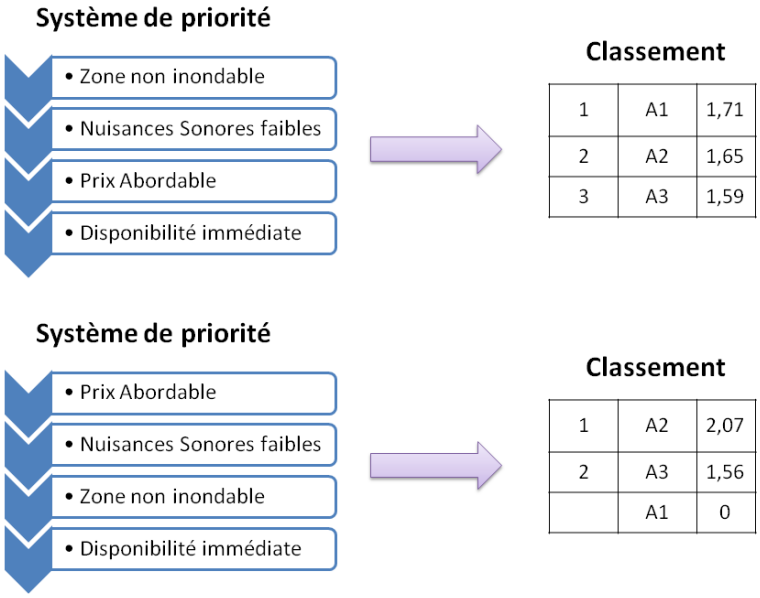


FIGURE 3.7 – Agrégation de Critères (avec l'approche par priorité)

à des couples d'attributs. Comme l'explique Labreuche et al. [LG03], il est nécessaire d'identifier les liens et de les quantifier, ce qui est difficile pour un utilisateur.

Dans le cadre de l'aide à la décision, l'agrégation de critères est utilisée afin de trouver les meilleurs compromis (si tous les critères ne peuvent être satisfaits) et donc aider un utilisateur à faire son choix. L'interaction avec l'utilisateur est importante et la plupart des travaux cités permettent à l'utilisateur de personnaliser l'agrégation. Ce dernier peut déterminer l'importance de chaque critère en lui affectant un poids, et moduler la compensation entre critères.

Après nous être intéressés à des approches de type combinaisons dans des domaines autres que la Recherche d'Information Géographique, nous allons maintenant présenter ce que proposent les systèmes de RIG.

3.4 Approches en Recherche d'Information Géographique

Dans la partie 2.5 nous avons présenté divers systèmes de Recherche d'Information Géographique existants. L'information géographique se compose de trois facettes (spatiale, temporelle et thématique), toutefois certains systèmes se sont intéressés uniquement à deux d'entre elles (généralement spatiale et thématique). Parmi les approches de RIG proposées, nous avons distingué quatre catégories : filtrage séquentiel, filtrage parallèle, combinaison linéaire, projection.

L'approche de filtrage séquentiel consiste à effectuer une recherche sur un critère, thématique par exemple, retournant un ensemble de résultats (documents pertinents

ordonnés). Ensuite une nouvelle recherche est lancée sur cet ensemble de résultats pour le deuxième critère. Puis de même pour le troisième. Les deux premières recherches permettent de réduire l'ensemble de documents résultats (approche de type filtre), et la dernière permet d'obtenir un ensemble de résultats ordonnés par rapport au dernier critère utilisé. Par exemple, si comme sur la figure 3.8 le dernier critère est le spatial, les documents seront ordonnés par degré de pertinence spatiale. Vu que chaque recherche (sauf la première) est effectuée sur un sous-ensemble de la base documentaire, cela allège et accélère les traitements. Cette approche de filtrage séquentiel permet de plus de ne pas avoir à combiner directement les trois facettes. Le système STEWARD [LSSS07], utilise une telle approche de filtrage séquentiel, mais uniquement sur le thématique et le spatial. Concernant l'ordre, Lieberman et al. [LSSS07] expliquent que l'idéal serait de déterminer l'ordre d'utilisation des critères (spatial puis thématique, ou, thématique puis spatial) selon la requête. En effet, si le critère spatial de la requête est très précis, il est plus intéressant de commencer par la recherche spatiale pour réduire au maximum le nombre de documents possibles pour la deuxième recherche.

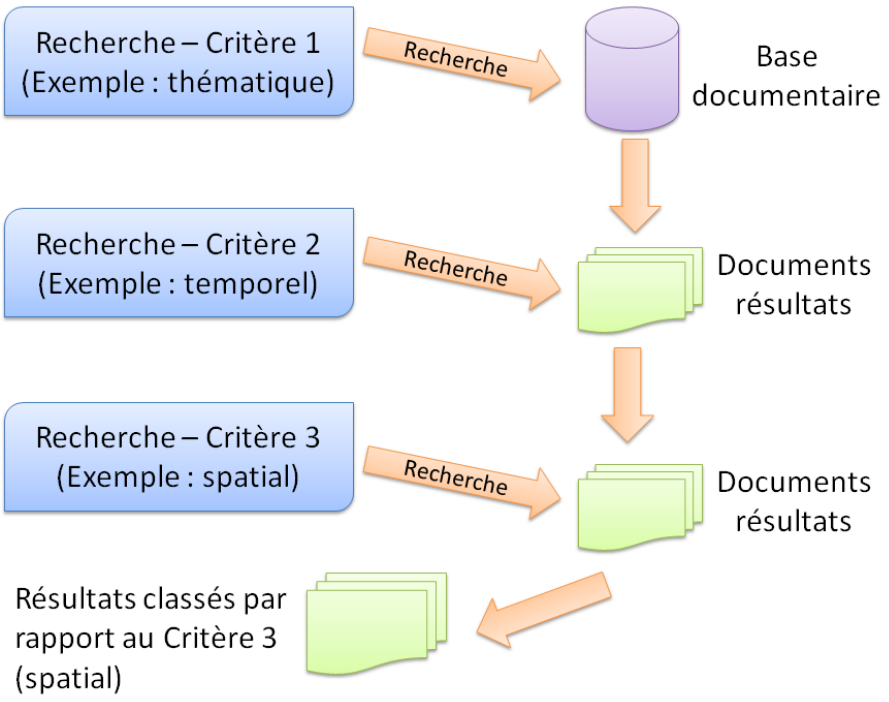


FIGURE 3.8 – Approche de filtrage séquentiel en RIG

L'approche de filtrage parallèle consiste à interroger séparément selon chacun des critères puis à réaliser l'intersection des résultats. Comme l'illustre la figure 3.9, la requête est soumise à trois traitements propres à chacun des critères géographiques. Pour chaque critère, un ensemble de résultats est retourné. Seuls les documents présents dans

les trois ensembles sont conservés. Par contre aucun classement n'est réalisé. Or dans l'ensemble de résultats obtenus, certains documents sont plus pertinents que d'autres, et donc devraient apparaître en tête de liste. Si cet ensemble de résultats est très important, les résultats les plus intéressants pourraient être placés très loin dans la liste. Cette approche de filtrage parallèle peut paraître nettement plus lourde en temps de traitement étant donné qu'elle réalise trois recherches complètes, néanmoins, comme son nom l'indique, l'indépendance de ces recherches rend la parallélisation possible. De plus, elle ne favorise pas un critère comme l'approche de filtrage séquentiel. Les systèmes SPIRIT [VJJS05], CITER [PEH+09] ainsi que WatWasWaar [LG08] utilisent une approche de filtrage parallèle.

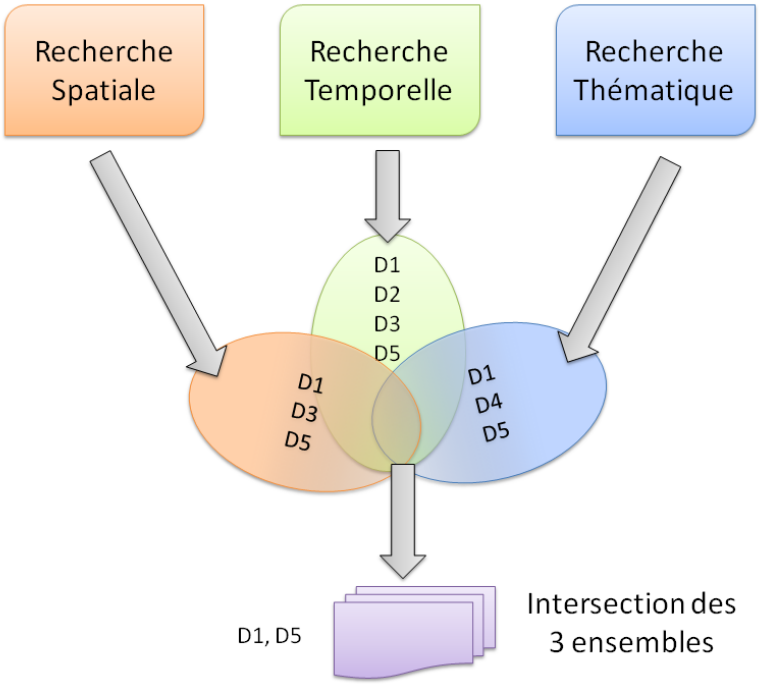


FIGURE 3.9 – Approche de type filtrage parallèle en RIG

Les approches de combinaisons linéaires permettent en RI de combiner des scores provenant de plusieurs systèmes [VC99], l'objectif étant d'obtenir un score unique. Généralement elles associent des poids aux scores obtenus par les différents systèmes. Par exemple la moyenne arithmétique est une combinaison linéaire avec une pondération équivalente pour chaque critère. La figure 3.10 illustre un exemple de combinaison linéaire sur deux SRI. Martins et al. [MSA05] proposent d'utiliser des approches de combinaisons linéaires, néanmoins ils ne précisent pas laquelle est utilisée dans DIGMAP [MMBS09]. GéoSem [BDEH07] utilise aussi une approche de combinaison linéaire : la moyenne arithmétique. Le système fait donc la moyenne des trois scores : spatial, temporel et

thématique. De même pour chaque critère, chaque SRI monofacette fait la moyenne des différents scores obtenus pour chaque document.

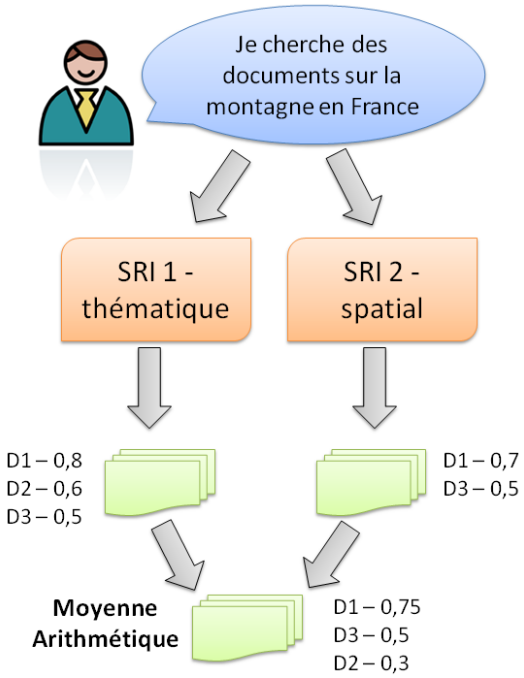


FIGURE 3.10 – Approche de combinaisons linéaires en RIG

Une dernière approche de combinaison, par projection, vise essentiellement la réorganisation des documents retournés lors d'une recherche pour améliorer leur diversité. Le but est de reléguer plus loin dans le classement des résultats trop proches pour augmenter la variété des résultats ou alors de les regrouper comme propose Google. L'approche propose de projeter les résultats sur un repère à n dimensions, les plus proches de l'origine du repère sont les plus pertinents. Ensuite les points du repère (les résultats) sont comparés par paire, et si un résultat est trop proche d'un autre il va être reclassé plus loin dans le classement. Van Kreveld et al. [VKRAVZ05] présentent l'approche sur deux facettes : spatial et termes. La figure 3.11 illustre un exemple sur ces deux facettes. Dans cet exemple, D2 est relégué à la fin du classement car trop proche de D1. Cette approche génère un classement ordonnée mais sans score, étant donné que certains résultats sont relégués plus loin dans le classement afin de proposer des réponses diversifiées en tête. Cette approche est justifiée par le fait que si la combinaison est réalisée en amont (moyenne arithmétique par exemple), il ne sera plus possible d'identifier les documents similaires. Par exemple, deux documents peuvent avoir un score de 0,5 pour des raisons différentes (D1 pour avoir 0,9 en spatial et 0,1 en thématique, D2 pour avoir 0,5 et 0,5). Cette approche empêche de moduler la combinaison (en favorisant un critère

par exemple), vu que le classement des résultats est modifié sans tenir compte de leur pertinence.

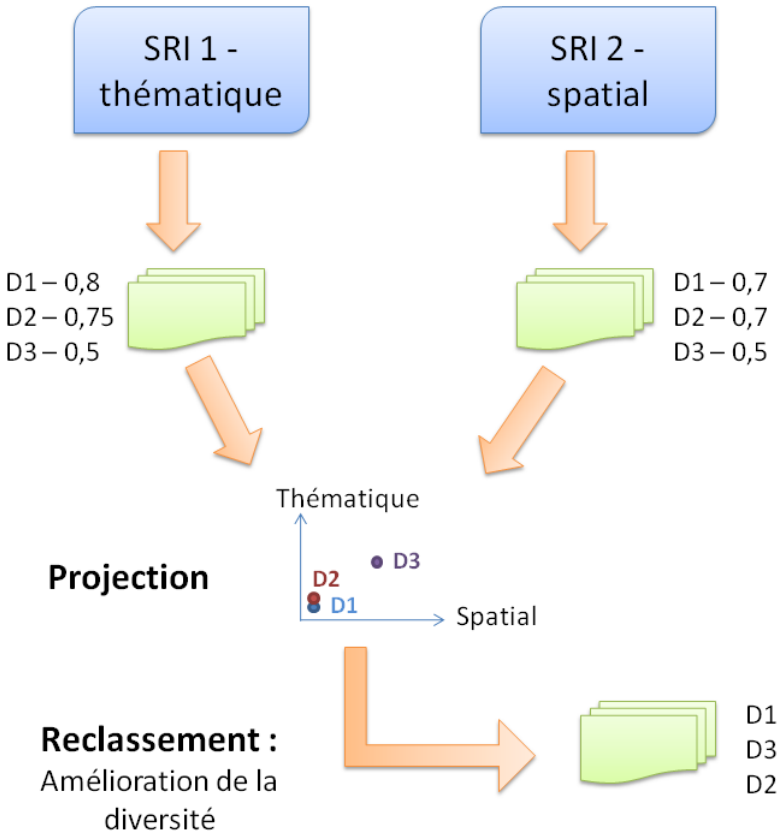


FIGURE 3.11 – Approche de type projection en RIG

Dans cette partie nous avons présenté quatre approches de combinaison de critères en Recherche d'Information Géographique : approche de filtrage séquentiel, approche de filtrage parallèle, approche de combinaison linéaire et projection.

3.5 Conclusion

Dans ce chapitre, nous avons présenté plusieurs approches de « combinaison » de critères : la fusion, l'agrégation, et les différentes approches utilisées en RI et RIG.

La fusion de critères, en Recherche d'Information Multimédia, permet de fusionner des informations provenant de différents types de documents (image et texte par exemple) ou de composants d'un document (pour une vidéo : les différentes images ainsi que la piste audio). Néanmoins, la fusion ne peut traiter que des représentations uniformisées

de ces informations. Par conséquent, ces informations doivent passer par une phase de généralisation (détaillée dans le chapitre suivant).

Un système d'aide à la décision nécessite une interaction importante. En effet, le système n'est pas censé trouver seul la solution, mais présenter un ensemble de propositions se rapprochant le plus des contraintes imposées par l'utilisateur. La plupart des approches proposent ainsi d'affecter des poids à chaque critère pour favoriser les plus importants ou pour régler le degré de compensation entre les critères.

En Recherche d'Information Géographique, nous pouvons distinguer quatre types d'approches. L'approche de **filtrage séquentiel** consiste à réaliser des recherches successives (une par critère) sur les ensembles de documents résultats obtenus après chaque recherche. Les premiers critères servent donc à filtrer (thématique et temporel par exemple), et le dernier à ordonner les résultats (spatial par exemple). L'approche de **filtrage parallèle** effectue une recherche sur chaque critère, puis, réalise l'intersection des ensembles de résultats obtenus pour ne conserver que ceux présents dans chacun des ensembles. Par contre ici, les résultats ne sont pas ordonnés. L'approche par **combinaison linéaire** permet de calculer un score unique à partir des différents scores (provenant de différents SRI). La dernière approche, par **projection**, vise surtout à détecter parmi les résultats ceux trop similaires pour les écarter et ainsi améliorer leur diversité. La plupart de ces approches sont des combinaisons indirectes (filtrage séquentiel ou parallèle, projection) ce qui permet d'éviter d'uniformiser les différents critères. Les approches de combinaisons linéaires permettent de combiner différents critères même s'ils n'ont pas été uniformisés, néanmoins cela peut introduire un biais comme nous allons le voir dans le chapitre suivant.

Dans ce chapitre, nous avons abordé les approches et techniques de combinaison. Mais certains travaux [MCF+03, PMLC07] soulignent le fait que lorsque les critères à combiner sont très différents, ils sont difficilement comparables. Dans certains cas, ils contournent le problème en ne les combinant pas directement (calcul sur les rangs par exemple [TWFM07]). D'autres travaux indiquent qu'une étape de « standardisation » est nécessaire [MCF+03]. Dans le chapitre qui suit nous allons nous intéresser à ces différentes approches d'uniformisation dans un but de combinaison de critères.

Chapitre 4

Uniformisation de critères

Sommaire

4.1 Introduction

Nous avons présenté dans le chapitre précédent différentes approches de combinaison de critères. Étant donné que les types de données à combiner sont généralement très hétérogènes, il est difficile de réaliser une telle combinaison. Certains travaux proposent des approches d'uniformisation pour résoudre ce problème. En Recherche d'Information Multimédia, la fusion de critères nécessite une phase de généralisation. L'aide à la décision multicritère a la particularité d'intégrer des critères pouvant être évalués qualitativement (par exemple : peu pertinent, pertinent, très pertinent). Pour combiner de tels critères avec des critères quantitatifs, il est nécessaire de les convertir via une approche de standardisation. En Recherche d'Information Géographique, comme nous l'avons vu dans le chapitre précédent, la plupart des approches proposent des techniques basiques de combinaison afin de s'affranchir des contraintes d'uniformisation. Ainsi ces approches impliquent des combinaisons relativement simples (pas de classement par exemple) ou peu personnalisables (pas de possibilité de favoriser certains critères comme en aide à la décision). Pour les autres approches, ne pas uniformiser peut introduire un biais dans le calcul du score global comme nous allons le voir dans la section suivante. Néanmoins, concernant l'information spatiale, il existe des approches de « focalisation/synthèse »

afin de n'associer qu'une information à chaque document ce qui revient à uniformiser le critère à l'extrême.

4.2 Normalisation en Recherche d'Information

La normalisation consiste à borner le score de pertinence d'un document calculé par rapport à une requête donnée. En recherche d'information, ces scores sont majoritairement bornés entre 0 et 1 [Kow97]. Si une requête est multicritère, chacun des critères doit être borné pour être comparable. Néanmoins, la normalisation ne garantit pas la comparabilité des scores obtenus sur 2 critères différents si les approches sont très spécifiques à chaque critère comme nous allons le voir dans cette partie.

Dans une recherche par mots-clés, un utilisateur peut spécifier plusieurs mots à trouver. La figure 4.1 illustre un tel exemple avec « vacances en montagne ». Un SRI classique va alors effectuer une recherche pour chaque terme, obtenir une liste de documents résultats ordonnés dont le score est normalisé (par exemple via la moyenne des scores des documents). L'unique ensemble de résultats obtenus est toujours associé à des scores normalisés. Étant donné que le même SRI est utilisé pour chacun des critères, il n'y a pas de problème de compatibilité entre les résultats obtenus pour chacun.

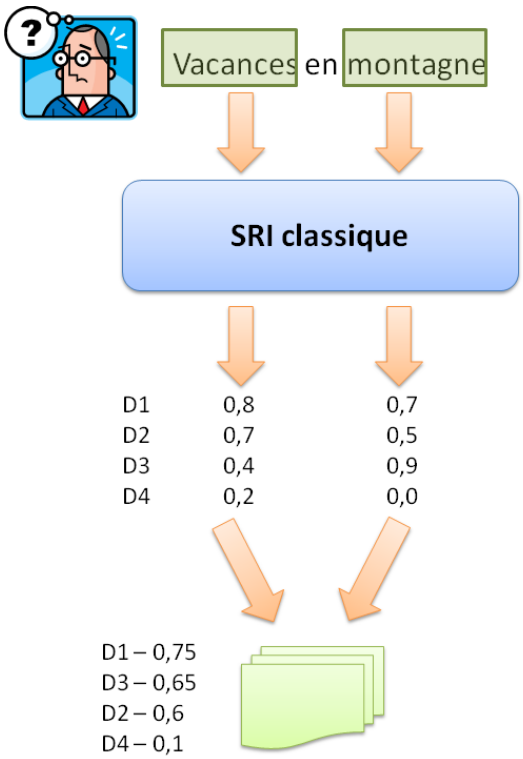


FIGURE 4.1 – Recherche d'information standard et normalisation

Prenons le cas de la Recherche d'Information Géographique et ses 3 facettes. Ici chaque facette est traitée via un SRI spécifique et adapté. Limitons nous ici à 2 facettes : le thème et le temporel. La figure 4.2 illustre un tel exemple avec « montagne en 1900 ». Pour le thème, en règle générale, les SRI géographiques utilisent un SRI standard (tel que Terrier) se limitant aux termes. Pour le temporel, un SRI dédié peut utiliser des approches très différentes de celles appliquées aux termes : pas de type unique (informations ponctuelles, périodes plus ou moins grandes), pas d'utilisation des fréquences d'apparitions, non prise en compte des autres informations temporelles du document, ... Cette différence peut introduire un biais, par exemple en surévaluant ou défavorisant le temporel par rapport au thématique. Par exemple, le système GéoSem [BDEH07] évalue une information temporelle de manière binaire (1 ou 0), puis, le score temporel du document est calculé via la moyenne les scores des informations temporelles qu'il contient. Néanmoins, si les documents du corpus contiennent rarement plus d'une information temporelle (comme les paragraphes dans notre corpus de type récits de voyage), le document aura donc souvent comme score 1, même si le document est peu pertinent temporellement (à cause de l'approche binaire). Or le thématique étant calculé via une approche statistique (TF·IDF) il aura rarement un score de 1 et est donc défavorisé par rapport au temporel. Dans la figure 4.2, même si D4 était en fait très peu pertinent temporellement (par exemple un score de 0,2), il se retrouve au même niveau que D2 qui est moyennement pertinent sur les 2 critères (D4 n'étant pertinent que pour le critère temporel).

La normalisation est donc une étape nécessaire à la combinaison (afin que les scores soient tous compris dans un même intervalle) mais non suffisante. En effet, si les scores sont obtenus via des approches hétérogènes, un biais peut être introduit comme nous l'avons souligné. Nous allons voir dans la suite de ce chapitre des approches pouvant compléter la normalisation pour résoudre ce problème.

4.3 Généralisation pour la Recherche d'Information Multimedia

La généralisation est une approche de simplification pour réduire la trop grande quantité de détails. Plusieurs approches permettent de réaliser cette simplification. La première est de type identification : à partir de données diverses et variées il s'agit de déterminer les concepts présents [Wol92] (un concept est une information porteuse de sens, une catégorie, tel que : voiture, homme, drapeau ...). Cela implique d'avoir au préalable défini l'ensemble des concepts et listé les données les caractérisant. La deuxième est de type regroupement/concentration : le but est de ne conserver que les informations les plus importantes d'un critère, regrouper les informations proches pour les mettre en avant. Dans certains travaux, notamment sur l'information spatiale, la hiérarchie des lieux (communes, départements, régions, ... par exemple) est utilisée pour regrouper les différentes informations spatiales [LSSS07, MMB08].

La généralisation de type identification est utilisée en Recherche d'Information Multimédia pour déterminer quels concepts contiennent un document. Il est nécessaire au

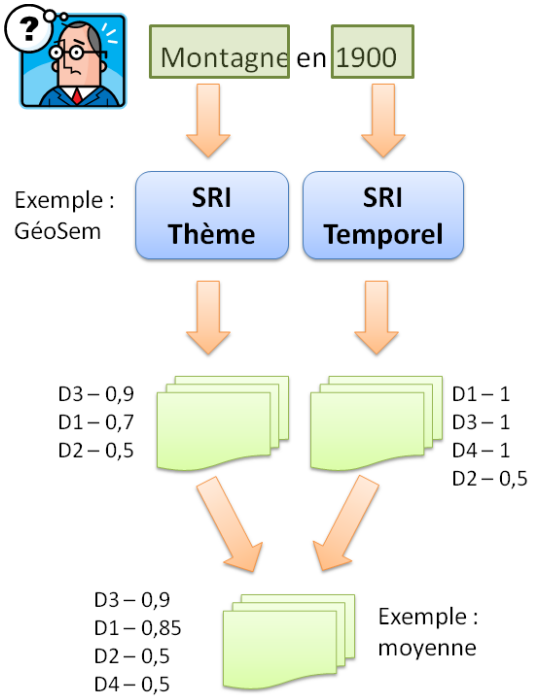


FIGURE 4.2 – Recherche d'information géographique et normalisation

préalable de définir la liste des concepts possibles et de les caractériser. De par sa richesse, il est difficile d'identifier directement les concepts contenus dans une vidéo. En effet une vidéo est un ensemble d'images se succédant rapidement (25 images par secondes) auxquelles peut s'ajouter une piste audio avec, dans certains cas, un discours. Il est donc possible d'appliquer l'approche de généralisation sur les différentes images de la vidéo et sur la piste audio pour identifier dans chaque cas les concepts présents. Ayache et al. [AGQ06] n'utilisent pas la même base de concepts pour les 2 types de documents (15 concepts pour les images, une centaine pour l'audio) étant donné que l'extraction d'information à partir d'image est très complexe et que certains concepts sont difficiles à identifier sur une image (exemple : politique). Néanmoins en utilisant la même approche de généralisation sur ces différents types de documents, il est possible de les fusionner comme expliqué dans la section 3.2.

La généralisation de type regroupement consiste à rassembler plusieurs informations (de même type) pour ne conserver que les plus importantes. Ce regroupement ne conduit pas nécessairement à des informations porteuses de sens comme pour la généralisation de type identification, ce qui permet de ne pas avoir à constituer une base de référence (tels que les concepts). Par exemple, la troncature en recherche d'information représente des termes par des termes tronqués (exemple : « avantages » et « avantager » sont regroupés en « avantag »). Pham et al. [PMLC07] proposent d'appliquer cette approche aux images : en découpant l'image avec une grille régulière, on obtient un ensemble de car-

reaux (appelés patches). Une image peut être ainsi considérée comme un sac de patches (tout comme un texte est considéré comme un sac de termes). Pham et al. [PMLC07] nomment ces patches des « visterms » (termes visuels) par analogie avec les termes contenus dans les textes (qu'ils nomment alors « texterms »). La figure 4.3 illustre le découpage d'une image en visterms via une grille régulière. Comme nous pouvons le voir, certains visterms apparaissent plusieurs fois dans l'image (tel que vt1 qui correspond au ciel bleu), donc il sera possible d'associer un poids à ces visterms par rapport à leur fréquence d'apparition. L'approche de segmentation de l'image en grille n'est pas la seule que Pham et al. [PMLC07] proposent. La segmentation par régions consiste à découper l'image en régions selon leurs contenus (couleurs par exemple) pour obtenir des visterms plus porteurs de sens. Néanmoins, cette approche ne permet pas d'identifier les régions détectées (par exemple l'image va être découpée en 6 régions, mais l'approche ne dit pas à quoi correspondent ces régions). Une fois l'ensemble de visterms obtenus, il est possible de leur appliquer un poids par rapport à leur fréquence d'apparition tout comme pour les termes. En vue de réaliser une combinaison avec une recherche textuelle classique, appliquer une approche similaire (troncature et modèle vectoriel) peut être une solution intéressante.

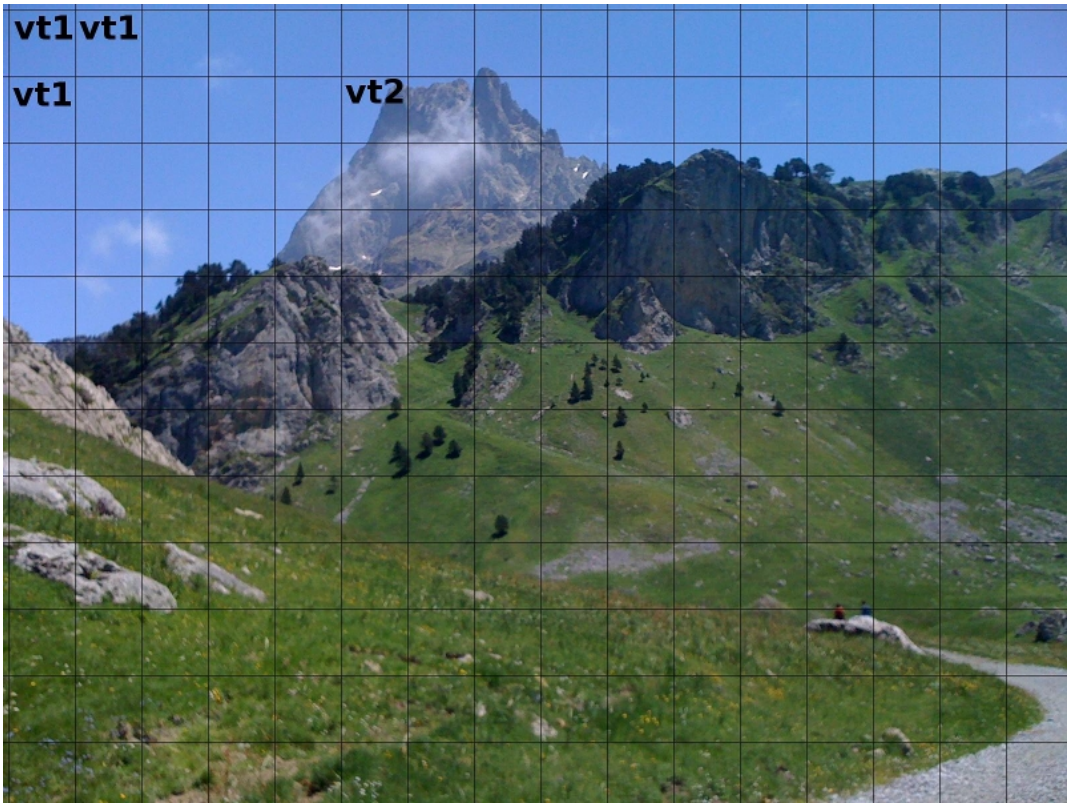


Figure 4.3 – Découpage d'une image en visterms

Nous avons donc présenté ici 2 approches de généralisation. La première de type identification cherche à déterminer quels concepts sont contenus dans un document. La deuxième de type regroupement réduit le nombre d'informations d'un critère pour ne garder que les plus importantes. La différence entre les deux approches est que la première traite des données de types différents contrairement à la deuxième.

4.4 Standardisation pour les Systèmes d'aide à la Décision

Utilisée notamment pour l'aide à la décision multicritère, la standardisation consiste, tout comme la normalisation, à ramener les scores des différents critères sur un même intervalle de résultats ([0;1] en général). Néanmoins, la différence importante avec la normalisation est que la standardisation s'applique à des évaluations non numériques [MCF+03, LG03]. Par exemple, dans [MCF+03], chaque critère est évalué selon 3 catégories (fort intérêt, intérêt moyen, faible intérêt) et dans [LG03] en 6 (très faible, faible, moyen, important, très important, exceptionnel). En plus de ces évaluations qualitatives, la standardisation nécessite, pour chaque critère, que les catégories soient comparées par paire. Dans [MCF+03], chaque paire est notée par l'utilisateur de 1 à 9 indiquant si les 2 catégories sont d'égales importances (1) ou si l'une est beaucoup plus importante que l'autre (9). Une fois cette matrice de comparaison de paires réalisée, le système génère les valeurs standardisées [Saa80, MCF+03]. La figure 4.4 illustre un exemple de standardisation. Un utilisateur recherche une voiture. Un des critères est la couleur. Il indique alors les couleurs idéales, celles qui sont acceptables et celles qui sont à la limite. Il doit de plus indiquer quel est l'écart entre 2 catégories (par exemple que l'écart entre la catégorie idéale et acceptable est faible). Le système peut ainsi générer un score pour chaque catégorie (par exemple 0,75 pour acceptable et donc les voitures vertes et oranges auront un poids de 0,75 pour la couleur).

La standardisation demande une paramétrisation avancée de la part de l'utilisateur. En effet, pour que le système puisse convertir les évaluations qualitatives en évaluations quantitatives, l'utilisateur est invité à classer les résultats possibles en degrés de satisfaisabilité (ou d'intérêt) et aussi à définir l'écart entre 2 degrés de satisfaisabilité en les comparant par paire. Cette approche demande donc un réglage par l'utilisateur qui peut s'avérer fastidieux. Néanmoins, étant donné que les systèmes d'aide à la décision doivent aider l'utilisateur selon ses préférences, ce type d'approche reste adapté. Nous allons maintenant aborder les approches utilisées en RIG.

4.5 La focalisation spatiale en Recherche d'Information Géographique

En RIG, chaque facette est traitée de manière spécifique. Chaque facette gère donc un type de données propre : pour le spatial des géométries, pour le temporel des périodes, et pour le thématique des concepts. Néanmoins, concernant le thématique, les systèmes de RIG utilisent souvent les approches de RI centrées sur les termes telles que présentées

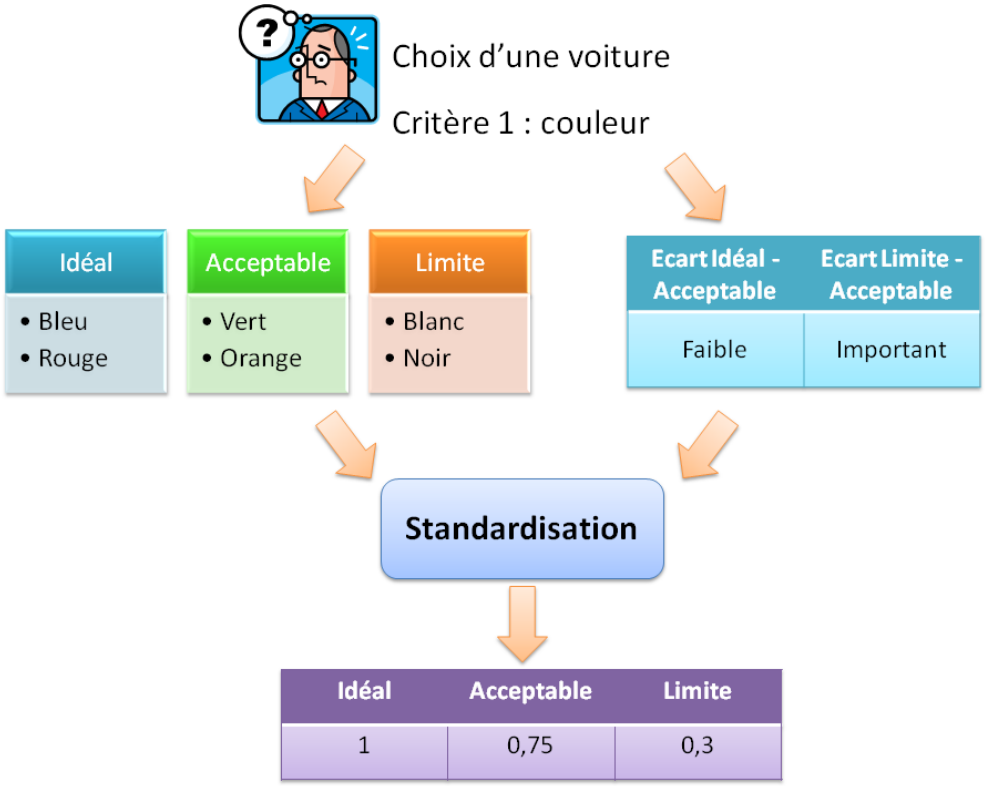


FIGURE 4.4 – Exemple de standardisation

dans la section 2.4.1.1. La plupart des systèmes ne traitant que les facettes spatiale et thématique, se pose alors la question de l'uniformisation d'information spatiale et de termes pour une combinaison des résultats.

Certains travaux [LSSS07, MMB08] proposent d'uniformiser les informations spatiales en cherchant notamment le « focus » du document, c'est à dire la zone principale à laquelle il se rapporte. L'avantage est que cela simplifie la recherche spatiale étant donné qu'une seule information spatiale est associée à chaque document. Les découpages administratifs étant hiérarchiques (le monde est divisé en continents, puis en pays, régions, villes, ...) et porteurs de sens, ils utilisent cette structure. Il est donc possible de regrouper plusieurs villes au sein d'une même région par exemple.

En appliquant une approche similaire au temporel et au thématique, c'est à dire en obtenant le contexte spatial, le contexte temporel et le contexte thématique du document, la combinaison s'appliquerait logiquement. En effet, chaque critère subissant une uniformisation de type généralisation, ils seraient comparables. Néanmoins, comme l'expliquent Lieberman et al. [LSSS07], il est très difficile d'obtenir un seul focus si le document contient des informations spatiales très éloignées, à moins de très fortement augmenter l'imprécision des résultats. Si, par exemple, un document parle de la ville de

Biarritz (au Sud-Ouest de la France) et de la ville de Lille (au nord de la France), la généralisation va retourner la France complète. La focalisation/synthèse, qui dans cet exemple conduit à une généralisation excessive, n'est donc pas recommandée [LSSS07].

4.6 Conclusion

Il existe différents types d'uniformisation selon les données utilisées et l'objectif visé. La normalisation consiste à borner des scores, généralement entre 0 et 1. Même si des scores sont bornés de manière équivalente, cela n'implique pas nécessairement qu'ils sont comparables. Certains peuvent être surnotés ou évalués différemment (de manière binaire par exemple) ce qui peut ajouter un biais s'ils sont couplés.

La standardisation permet de convertir des évaluations qualitatives (non numériques) en évaluations quantitatives (sur l'intervalle [0;1]). Elle s'applique en aide à la décision multicritère. Elle demande certains paramétrages de l'utilisateur pour que le système soit capable de faire cette conversion. Tout comme la normalisation, les scores sont bornés entre 0 et 1.

La généralisation permet de ne conserver que les informations les plus importantes. Il existe 2 types de généralisations. La première vise l'identification des concepts (ou catégories telles que : voiture, homme, ...) présents dans un document à partir d'informations de différents types (par exemple extraites d'images et de textes). La deuxième regroupe des informations de même type pour ne conserver que les plus importantes. Ce regroupement implique aussi que toutes les informations soient rattachées à un seul type (par exemple les *visterms* pour une image). Cette approche permet de plus d'utiliser des méthodes statistiques telles que la fréquence d'apparition éprouvée en RI. Ainsi il est possible de la combiner directement avec l'approche utilisée pour les termes. La généralisation implique néanmoins une perte d'information puisque elle consiste à éliminer les détails.

En recherche d'information géographique, l'approche de focalisation, qui s'assimile à une généralisation maximum, cherche à n'associer qu'une information spatiale à un document. Cette approche pose problème lorsque les informations spatiales sont très éloignées.

Conclusion de l'état de l'art

Nous avons choisi de traiter chaque facette de l'information géographique de manière spécifique et indépendante comme le préconisent de nombreux travaux en RIG tels que [CJP06, MSA05]. Afin de permettre aux utilisateurs de faire des recherches sur plusieurs facettes il est nécessaire de les combiner. Notre problématique principale est donc de trouver comment combiner les résultats issus de chaque SRI dédié (spatial, temporel et thématique) lors de la phase d'interrogation du corpus.

Le chapitre 3 a présenté différentes approches de combinaison de critères en Recherche d'Information et dans d'autres domaines. Comme nous avons pu le constater, celles utilisées en RIG se limitent essentiellement soit à une approche de type filtrage en réalisant l'intersection des ensembles de résultats (donc il n'y a pas de classement de résultat), soit à des méthodes linéaires ne permettant aucun paramétrage de la combinaison. En ce qui concerne les approches d'aide à la décision multicritère, elles proposent, pour chaque critère, de spécifier son importance ou encore de régler le degré de compensation entre les critères. Cette flexibilité nous intéresse particulièrement et nous avons donc décidé d'adapter ces approches pour notre proposition de combinaison par contrainte étendue.

Le chapitre 4 a présenté différentes approches d'uniformisation. En recherche d'information géographique, certains systèmes se sont passés d'uniformisation en utilisant des approches de combinaisons très simples (tel que le filtrage). Dans les autres cas les combinaisons, notamment via des approches linéaires, ont été effectuées sans uniformisation. Comme nous avons pu le constater, cela peut introduire des biais (critère avantagé). Étant donné que chaque facette de l'information géographique est traitée avec un SRI dédié (un spatial, un temporel et un thématique) (voir chapitre 2), nous avons donc trois approches d'indexation et trois approches de RI différentes. L'approche proposée par Pham et al. [PMLC07] pour traiter les images de manière similaire aux textes, c'est-à-dire l'utilisation de termes visuels (*visterms*) et le calcul d'un poids par rapport à la fréquence d'apparition, nous intéresse particulièrement. En effet, les approches statistiques appliquées aux termes (TF·IDF, modèle vectoriel de Salton) sont utilisées depuis des années et donnent de bons résultats [BYRN99]. Nous souhaitons donc adapter de telles approches statistiques au traitement des informations spatiales et temporelles, afin d'utiliser une approche homogène pour les différentes facettes de l'information géographique. Il s'agit, d'une part, d'homogénéiser les formes de représentation de l'information spatiale et de l'information temporelle et, d'autre part, de mettre en œuvre une même approche statistique de calcul de score de pertinence.

Le chapitre 2 a détaillé les traitements impliqués dans la Recherche d'Information Géographique textuelle. Comme nous pouvons le voir, il existe divers SRIG traitant seulement certaines facettes de l'information géographique (généralement spatiale et thématique) ou les trois pour certains. Néanmoins il n'existe pas de cadre d'évaluation permettant d'évaluer un SRIG. Afin de tester et d'évaluer des SRIG supportant la combinaison des trois facettes, il est donc nécessaire de mettre en place un tel cadre. Pour les SRI textuels il existe des campagnes d'évaluations reconnues tel que TREC. Nous avons donc décidé de capitaliser ce savoir-faire en y intégrant les spécificités relatives à l'information géographique.

Troisième partie

Contribution : vers la combinaison par contraintes de critères de recherche en RIG

Introduction de la contribution

Dans la partie précédente nous avons présenté l'état de l'art lié au traitement automatique de l'information géographique ainsi qu'à la combinaison et l'uniformisation de critères. Notre objectif est la combinaison de critères géographiques (spatiaux, temporels et thématiques). Dans cette partie nous allons présenter les différentes contributions proposées pour arriver à cet objectif.

La première est une approche d'uniformisation générique que nous appliquons à l'information spatiale et à l'information temporelle. Il s'agit d'adapter une stratégie appliquée en RI classique sur les termes, tout comme Pham et al [PMLC07] l'ont fait pour les images (voir section 4.3). Cette uniformisation nous permet d'envisager des combinaisons plus avancées que celles proposées en RIG actuellement (voir section 3.4). Cette proposition est présentée dans le chapitre 5.

La deuxième contribution consiste à évaluer la combinaison linéaire classique de différentes facettes de l'information géographique et à quantifier l'apport de cette combinaison. Pour cela, nous proposons d'utiliser des approches linéaires standards ayant fait leurs preuves en RI classique. Ces approches sont présentées dans le chapitre 6 et l'évaluation est présentée dans le chapitre 8.

La troisième contribution est une approche de combinaison, originale et générique, basée sur les contraintes que nous appliquons à la RIG. Cette approche s'inspire de celles utilisées habituellement pour l'aide à la décision (voir section 3.3) permettant à un utilisateur de personnaliser la manière dont sont combinés différents critères. Cette approche est présentée dans le chapitre 6.

La dernière contribution propose un cadre expérimental permettant d'évaluer un SRIG. Etant donné qu'il n'existe pas de tel cadre (voir section 2.4.4), pour évaluer les différentes approches de combinaisons présentées, nous avons défini et mis en place ce cadre expérimental en adaptant celui utilisé dans les campagnes TREC (voir section 2.4.4). Il est présenté dans le chapitre 6.

Cette partie Contribution s'articule ainsi sous la forme de quatre chapitres. Le premier (chapitre 5) présente notre première contribution portant sur l'uniformisation. Le deuxième (chapitre 6) présente les deux contributions portant sur la combinaison de critères ainsi que sur l'évaluation d'un SRIG. Le troisième (chapitre 7) présente les différents prototypes mis au point dans le cadre de ce travail. Pour finir le dernier chapitre de cette partie (chapitre 8) présente toutes les expérimentations réalisées ainsi que leurs analyses.

Chapitre 5

Uniformisation de données

Sommaire

5.1 Introduction

Comme nous avons pu le voir dans le chapitre 3, l'hétérogénéité des données contenues dans certains documents (par exemple une vidéo) ou informations (par exemple les informations géographique) nécessite leur décomposition en plusieurs critères (images et bande sonore pour une vidéo ; trois facettes pour l'information géographique) et donc la combinaison par la suite. En ce qui concerne l'information géographique, les différentes facettes (spatial, temporel et thématique) peuvent être traitées par un SRI classique comme c'est le cas aujourd'hui mais de manière limitée. Un SRI non géographique n'est pas capable d'interpréter des relations spatiales ou temporelles telle que « près de ». Il est donc nécessaire de traiter de manière spécifique chaque facette de l'information géographique et de les combiner par la suite. Avant de réaliser cette combinaison, il est important d'uniformiser les représentations des données ainsi que les démarches de traitement de ces données relatives aux différentes facettes afin d'éviter des biais comme expliqué dans le chapitre 4. Notre objectif est donc d'utiliser une approche générique d'uniformisation applicable sur chacune de ces facettes.

Les méthodes classiques d'indexation automatiques [BYRN99] procèdent à la normalisation des termes considérés comme significatifs dans le corpus, puis, calculent un poids correspondant à la fréquence d'apparition de chacun de ces termes. Nous proposons donc d'appliquer une approche similaire à chaque critère à uniformiser : regrouper les informations en un seul type, que l'on peut qualifier de pivot (tel que les lemmes pour les termes), puis calculer leur fréquence d'apparition. Cette approche s'inspire aussi de celle proposée par Pham et al. [PMLC07] qui ont montré qu'il était possible d'adapter ces méthodes classiques de RI sur les images.

Ce chapitre présente donc notre première contribution qui consiste en une approche d'uniformisation générique que nous appliquons aux informations spatiales et temporelles.

5.2 Indexation multidimensionnelle basée sur le « tuilage »

Le « tuilage » est une segmentation en « tuiles ». Étant donné que les éléments obtenus via un découpage ne sont pas nécessairement de forme carrée ou rectangulaire (carreaux) nous avons choisi de les nommer « tuiles ».

L'approche que nous proposons consiste à uniformiser les représentations des données indexées, ou objets, afin de générer de nouveaux index puis de combiner les résultats obtenus lors de leurs interrogations. Il est donc nécessaire de redécouper l'espace représentant l'ensemble des objets en un ensemble de sous-espace : les tuiles. Chaque objet est alors rattachée à une ou plusieurs tuiles, tout comme les termes avec les lemmes (mis à part qu'un terme n'est rattaché qu'à un lemme). À ces tuiles peut ensuite être associé un poids basé sur leur fréquence d'apparition. Notre approche se base donc sur des index préexistants qu'il s'agit d'uniformiser.

Selon le type de tuilage souhaité (voir section 5.2.3), la représentation uniforme qu'est la tuile peut se baser sur deux approches. La première utilise un découpage géométrique paramétré selon l'unité souhaitée pour générer le tuilage. La deuxième est un découpage « explicite » qui est bâti sur le sens commun tel qu'un découpage administratif. Cette approche est plus difficile à formaliser car elle peut se baser sur plusieurs paramètres.

Selon les critères, l'information peut être représentée sur une ou plusieurs dimensions. Par exemple l'information temporelle est représentée sur une dimension, alors que l'information spatiale peut être représentée sur deux (planaire) ou trois (avec la hauteur de chaque point) dimensions.

L'indexation est constituée de deux étapes : la mise en place du tuilage et la pondération des tuiles. Ainsi, la première étape consiste à générer des tuilages de grains différents. La deuxième consiste à assigner à chaque tuile un poids en fonction de sa fréquence d'apparition dans un document. À l'issue de ces deux étapes, nous disposons donc de plusieurs nouveaux index uniformisés.

5.2.1 Approche de tuilage

L'approche de tuilage consiste à appliquer aux représentations initiales un tuilage à autant de dimensions que comportent ces représentations comme l'illustre la figure 5.1. Cette approche est une forme de discrétisation. 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. L'approche utilise donc un index existant (toutes les données à uniformiser sont dans cet index) et produit un nouvel index contenant les données uniformisées.

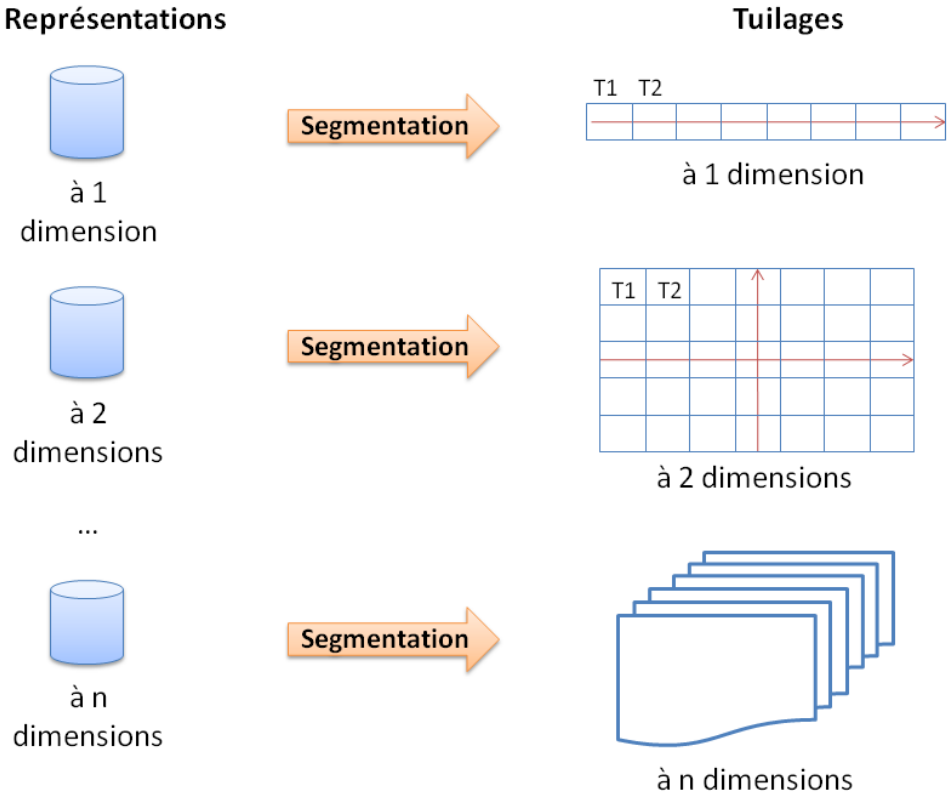


FIGURE 5.1 – Approche de tuilage

Nous allons maintenant décrire cette uniformisation de manière formelle. Au domaine [31] O inclus dans l'espace R^n [32] 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

31. 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 ...

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

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}).

$$\begin{aligned} O \subseteq \mathbb{R}^n \longrightarrow T \subseteq \mathbb{R}^n \\ O = \{O_1, O_2, O_3, \ldots, O_p\} \\ T = \bigcup_{i=1}^{m} T_i \\ N_{T_i} = \mid T_i \mid T_i \cap O_j \neq \varnothing \quad \forall j = 1, \ldots, p \mid \\ \text{avec } \mid x \mid \text{ la cardinalité de } x \end{aligned} \tag{5.1}$$

Prenons le cas de l'information spatiale pour illustrer cette approche de tuilage. La figure 5.2 illustre le contenu d'un index spatial : quatre représentations (trois communes : Gèdre, Cauterêts et Barèges ; un pic : Vignemale ; une relation spatiale : près de Gavarnie). Un tuilage est alors généré (ou choisit si nous disposons de tuilages existants) pour couvrir l'ensemble des représentations contenues dans l'index initial tel que sur la figure 5.3. Puis l'intersection de ces représentations et du tuilage permet d'obtenir l'ensemble des tuiles correspondant au corpus (figure 5.4).

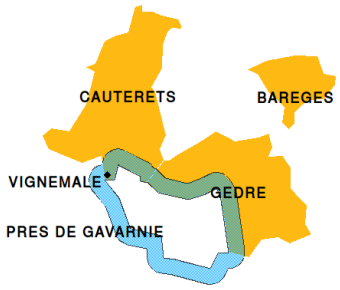


FIGURE 5.2 – Représentations spatiales

FIGURE 5.3 – Tuilage généré par rapport aux représentations existantes

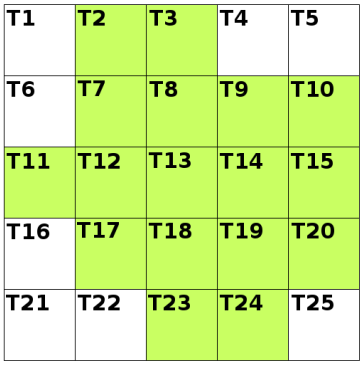


FIGURE 5.4 – Tuilage conservé (tuiles colorées)

Dans les travaux présentés dans l'état de l'art, l'approche des *visterm* [PMLC07] (voir section 4.3) propose un découpage des images. Dans cette approche, chaque partie

d'image est associée à un *visterm*, c'est à dire des « termes visuels », afin de reproduire les approches utilisées pour les termes contenus dans des textes en RI. Dans ce cas, un morceau d'image est rattaché à un seul *visterm*. Notre approche se démarque de celle-ci en permettant d'uniformiser des objets de diverses dimensions (1 à n) mais aussi de différentes échelles (une représentation peut impacter plusieurs tuiles).

Étant donné que les facettes de l'information géographique que nous étudions, le temporel et le spatial (cartographique), sont représentées respectivement sur une et deux dimensions, nous nous sommes plus particulièrement intéressés à ces cas. Ils seront explicités et illustrés dans la section 5.2.4 qui va suivre.

5.2.2 Tuilage multi-échelle

Dans certains cas, l'information peut être représentée via différentes échelles aussi bien dans les requêtes des utilisateurs que dans les documents à indexer. Lors de l'indexation des documents, il est donc possible de générer des index adaptés à chaque échelle. Ainsi, comme l'illustre la figure 5.5, selon l'échelle de la requête, le SRI va pouvoir sélectionner l'index le plus adapté. Par exemple, si la requête concerne la région « Aquitaine », alors le tuilage de type région sera utilisé.

Néanmoins, il n'est pas toujours possible de déterminer sur quelle échelle porte la requête : par exemple si cette requête porte sur plusieurs informations d'échelles différentes. En effet, si la requête concerne une commune et un département (par exemple : « à Pau et dans les Hautes Pyrénées »), quel tuilage alors choisir ? Dans ce cas, il est nécessaire d'avoir un tuilage « par défaut » qui serait utilisé lorsque l'échelle de la requête n'est pas identifiable. Ce tuilage est alors à sélectionner par rapport au contenu du corpus : par exemple en réalisant une étude statistique de l'échelle moyenne des informations présentes dans le corpus. Comme nous allons le voir dans le chapitre 8, le type d'information contenu dans le corpus est prédominant pour le choix du tuilage.

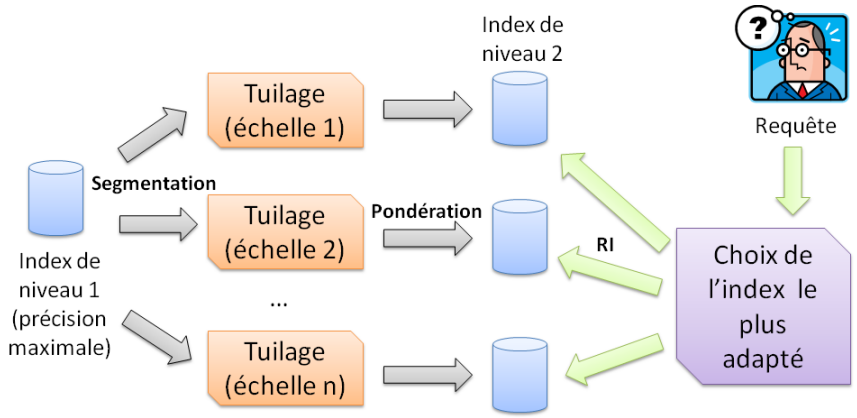


FIGURE 5.5 – Indexation multi-échelles

5.2.3 Types de tuilages

Outre le tuilage multi-échelle, il est possible d'utiliser des tuilages de différents types. Nous avons identifié deux types de tuilages : le tuilage « régulier » et le tuilage « explicite ».

Le tuilage « régulier » est un tuilage consistant à découper la zone couverte par l'ensemble des objets en tuiles de taille similaire sans tenir compte de ces informations. Cette approche est semblable à la troncature. Ce tuilage se base donc sur un découpage géométrique en spécifiant une unité de référence (telle que la distance pour le spatial). L'avantage de ce tuilage est que la taille des tuiles et leurs frontières sont réglables. Il est ainsi défini par rapport à un corpus donné. Néanmoins à chaque modification de ce dernier, il est possible que le tuilage change (si par exemple une nouvelle information indexée se situe en dehors de la zone de référence). Ce type de tuilage est présenté par Pham et al. [PMLC07] pour les *visterms*.

Le tuilage « explicite » consiste à utiliser un tuilage déjà défini et porteur de sens. Cette approche est semblable à la lemmatisation. Donc contrairement au tuilage régulier, ici nous ne générons pas de tuilage. De plus, à ces tuiles sont associées des informations supplémentaires (par exemple, nom : Pau ; échelle : commune, ...). Ce tuilage se base donc sur un découpage que nous avons qualifié de « signifiant » car bâti sur des critères humains (de sens commun). L'avantage est donc que ces tuiles sont fixées et identifiées. Un autre avantage pour les représentations à plus de deux dimensions est que ce tuilage peut être plus précis (par exemple pour la 2D des polygones) étant donné que les représentations ne sont pas limitées à une seule forme comme avec le tuilage régulier. Le principal inconvénient qui découle de cette approche est inhérent au caractère fixe des tuiles : deux informations limitrophes (et donc proches) peuvent être rattachées à deux tuiles différentes alors que nous pourrions souhaiter les voir rattachées à la même tuile.

Il existe donc divers types de tuilages, chacun ayant leurs avantages et inconvénients. Le tuilage régulier peut s'appliquer à tout type d'information. Par contre, le tuilage explicite nécessite donc de posséder un découpage porteur de sens. Il est possible que pour certains types d'information on ne dispose pas de ce type de tuilage ou qu'il n'existe pas. Nous allons maintenant illustrer ces différents types de tuilages par des propositions dédiés à l'information géographique.

5.2.4 Application à l'information géographique

Pour l'information temporelle, deux types de tuilages sont possibles : le tuilage régulier (figure 5.7) et le tuilage calendaire (figure 5.6). Le tuilage calendaire est un tuilage de type « explicite » et donc consiste à utiliser le découpage standard : jours, semaines, mois, saisons, années, siècles, ... Ce découpage permet la définition de plusieurs index correspondant à ces différents niveaux de précision (mois, jours, ...).

De même pour l'information spatiale, deux types de tuilages sont possibles : tuilage administratif (figure 5.8) et tuilage régulier (figure 5.9). Le tuilage administratif est un tuilage de type « explicite » et donc consiste à utiliser un découpage standard : quartiers, communes, cantons, départements, régions, pays, ... Ce découpage permet la définition

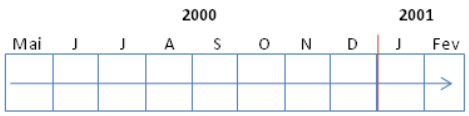


FIGURE 5.6 – Tuilage calendaire (Mois)

FIGURE 5.7 – Tuilage régulier (Tuiles de 40 jours)

de plusieurs index correspondant à ces différents niveaux de précision (communes, départements, ...).

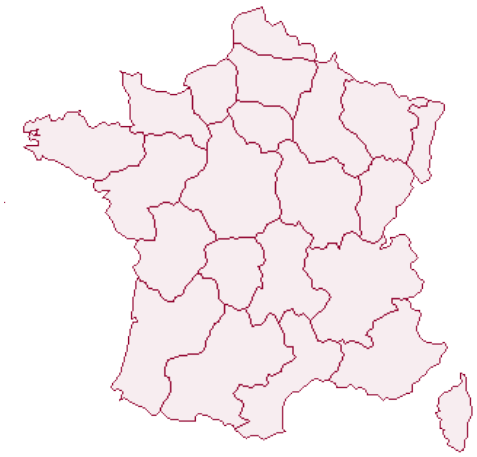


FIGURE 5.8 – Tuilage administratif (Régional)

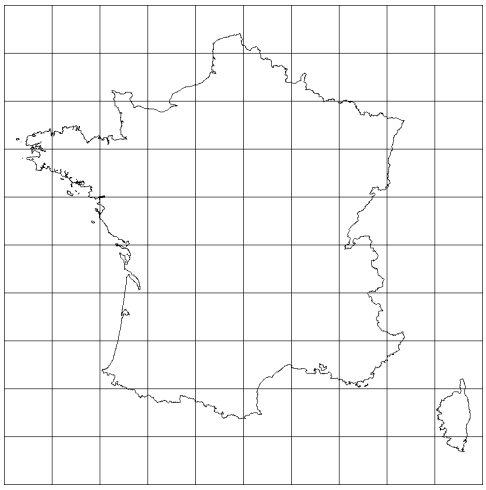


FIGURE 5.9 – Tuilage régulier (10x10)

Nous avons présenté les différentes approches de tuilage permettant d'uniformiser les objets en les projetant sur l'ensemble de tuiles générées. Une fois définies, ces tuiles, similaires aux lemmes, vont être pondérées, comme nous allons le voir maintenant.

5.2.5 Pondération des tuiles

Afin de déterminer l'importance de chaque tuile dans un document, un poids leur est associé. Ce poids est basé sur la fréquence d'apparition de la tuile. Nous proposons deux approches discrètes : une fréquence binaire et une fréquence proportionnelle (voir tableau 5.1). Comme nous allons le présenter à la fin de cette section, ces deux fréquences seront normalisées.

La fréquence binaire a pour valeur : 1 s'il y a intersection entre l'objet et la tuile, 0 sinon. Donc chaque objet intersectant la tuile incrémente sa fréquence de 1. L'approche permet ainsi de traiter chaque objet de manière équivalente : que ce soit un point, une ligne ou une surface, l'incrémentation est la même. Néanmoins, un objet dont la représentation n'intersecte par exemple que 10 % de la tuile impacte autant qu'un objet

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

TABLE 5.1 – 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)

correspondant à la tuile complète. Dans le premier cas, la fréquence de la tuile ne devrait-elle pas être incrémentée uniquement de 10 % de la valeur maximum (soit de 0,1) ?

Prenons trois exemples :

1. un document contient une information représentée sous la forme d'une représentation ponctuelle. Dans ce cas là, l'approche binaire est plus adaptée.
2. un document contient deux informations représentées sous la forme de surfaces de deux échelles différentes (par exemple un jour et un mois). Si le tuilage utilisé est d'échelle supérieure à ces deux là (par exemple année), il est plus intéressant que celui des deux qui a l'échelle la plus proche du tuilage donne un poids plus important à la tuile (c'est-à-dire le mois). Donc l'approche proportionnelle est plus adaptée.
3. un document contient deux informations, l'une sous la forme d'une représentation surfacique, l'autre représentée sous la forme d'une représentation ponctuelle incluse dans la première. Admettons que la représentation surfacique correspond exactement à une tuile. Les deux informations ne devraient pas donner le même poids. L'approche proportionnelle est la plus adaptée. Toutefois pour l'objet ponctuel, son poids risque d'être négligeable. Une pondération minimale permettrait de ne pas trop pénaliser ces objets.

Pour tenir compte de cette particularité, nous proposons une fréquence proportionnelle. Selon le ratio de recouvrement objet/tuile (dans la formule : surface de l'information sur surface de la tuile), la fréquence associée à la tuile est incrémentée d'une valeur plus ou moins importante comprise entre 0 et 1. Nous avons pondéré la fréquence de chaque information par le nombre de tuiles couvertes par cette information. Le but est de réduire le poids des représentations très grandes.

Reprenons l'exemple de la section 5.2.1. La figure 5.10 illustre un tuilage régulier sur des objets spatiaux (O_1 à O_5). Le tableau 5.11 illustre les pondérations binaires et proportionnelles obtenues par rapport à ce tuilage. Comme nous pouvons le voir, la tuile T2 est intersectée par une objet (O_1) et donc sa fréquence binaire est de 1. Si nous tenons compte du ratio de recouvrement entre les objets et la tuile, sa fréquence proportionnelle n'est que de 0,10. Pour la tuile T12, trois objets l'intersecte donc sa fréquence binaire est de 3 (mais 0,75 pour la fréquence proportionnelle).

Nous exploitons ces fréquences pour le calcul du poids associé à chaque tuile. Nous avons utilisé quatre formules de pondérations. Les trois premières (TF, TF·IDF, OkapiBM25) sont très utilisées en RI. Elles sont appliquées à des fréquences binaires nor-

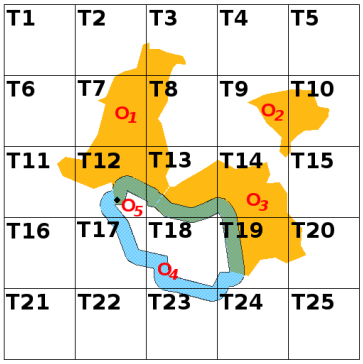


FIGURE 5.10 – Tuilage régulier sur des objets spatiaux

id_t	liste id_{O_i}	fréquence binaire	fréquence proportionnelle
T1	[/]	0	0
T2	[O_1]	1	0,10
...			
T12	[O_1,O_4,O_5]	3	0,75
...			

FIGURE 5.11 – Exemple d'index par rapport au tuilage régulier de la figure 5.10

malisées. Au lieu de les appliquer sur les termes, nous proposons de les utiliser sur les tuiles. L'IDF réduisant le poids des tuiles fréquentes, nous avons décidé d'utiliser le TF seul aussi afin de comparer les deux approches. Pour tester la fréquence proportionnelle, nous avons choisi le TF car cette formule permet notamment de ne pas réduire le poids des tuiles fréquentes. Nous avons appelé cette formule TFp (pour TF proportionnel). Le tableau 5.2 présente ces différentes formules que nous proposons d'expérimenter. Il faut noter que ces différentes formules normalisent les fréquences calculées auparavant : pour le TF et TFp en divisant, pour un document donné, le score par le nombre de tuiles contenus dans ce document.

À la fin de ce processus d'indexation, nous avons donc de nouveaux index. Ces index uniformisés contiennent pour chaque tuile son poids, l'identifiant du document et du paragraphe correspondant. Une fois ces index obtenus, il est possible d'appliquer des stratégies de Recherche d'Information permettant de tirer avantage des tuiles.

5.3 Approches de recherche d'information appliquées au tuilage

Tout comme l'approche d'indexation basée sur les tuiles imite l'approche standard utilisée pour les termes, nous souhaitons appliquer les stratégies de RI utilisées couramment sur ces tuiles.

Le modèle vectoriel de Salton [Sal71] (voir section 2.4.3 page 33) est très utilisé en RI [Bes04]. Nous avons donc décidé de l'appliquer aux tuiles : cela consiste à représenter l'ensemble des tuiles d'un document sous la forme d'un premier vecteur ainsi que l'ensemble des tuiles de la requête sous la forme d'un deuxième vecteur et de comparer ces deux vecteurs. Cela revient à représenter la base documentaire sous la forme d'une matrice, comme sur la figure 5.3 (D correspond à un document, T à une tuile, et w_{ij} au poids de la tuile j pour le document i). Le score d'un document sera, par exemple, calculé en utilisant le produit scalaire (voir section 2.4.3.1).

Tile Frequency (TF)	$W_{T_i,Du} = TF(T_i, Du) = \frac{freq(T_i,Du)}{\sum_{i=1}^{n} freq(T_i)}$
TF·IDF	$W(T_i, Du) = TF(T_i, Du) * IDF(T_i) \quad avec\ IDF(T_i) = log\left(\frac{NDu}{NDu_{T_i}}\right)$
OkapiBM25	$W(T_i, Du) = \left(\frac{(k_1+1)*TF(T_i,Du)}{(K+TF(T_i,Du))}\right) \quad avec\ K = k_1 * [(1-b) + \frac{b*n}{advl}]$
TFp	$W(T_i, Du) = TFp(T_i, Du) = \frac{freqP(T_i,Du)}{\sum_{i=1}^{n} freq(T_i)}$
$freq(T_i, Du)$: fréquence de la tuile T_i dans l'unité documentaire Du, $freqP(T_i, Du)$: fréquence proportionnelle de la tuile T_i dans l'unité documentaire Du, n : nombre de tuiles dans l'unité documentaire Du, NDu_{T_i} : nombre d'unités documentaires contenant la tuile T_i, NDu : nombre d'unités documentaires, $k_1 = 1{,}2$, $b = 0{,}75$, $advl = 900$	

TABLE 5.2 – Formules de pondération appliquées aux index uniformisés, pour une tuile T_i et une unité documentaire Du

5.4 Conclusion

Notre hypothèse est que la combinaison des différentes facettes de l'information géographique améliore la pertinence des résultats. Néanmoins il est nécessaire d'homogénéiser les données relatives à ces facettes dans de nouveaux index avant de les combiner. Nous avons donc proposé dans ce chapitre une approche d'uniformisation générique pouvant être appliquée à l'information spatiale et à l'information temporelle.

Cette approche consiste à uniformiser, selon une signification (sens commun) ou une mesure, les données d'index dédiés aux facettes spatiales et temporelles afin de combiner les résultats obtenus lors de leur interrogation. Il s'agit de générer de nouveaux index basés sur un seul type de représentation : la tuile. Cette indexation correspond à deux étapes. La première étape est le choix du tuilage. Un ensemble d'objets à une dimension (telle que le temporel) sera découpé selon un axe à une dimension (telle qu'une ligne de temps) et un ensemble d'objets à deux dimensions sera découpé selon un repère à deux dimensions (telle qu'un système à coordonnées planaires). Plusieurs types de tuilages sont possibles : régulier ou explicite (calendaire pour le temporel, administratif pour le spatial). La deuxième étape consiste à pondérer ces tuiles par rapport à leur fréquence d'apparition. Cette dernière peut être soit binaire (1 si intersection, 0 sinon), soit proportionnelle (valeur entre 0 et 1 selon le degré d'intersection). Pour le calcul de pondération, nous proposons d'utiliser les formules utilisées couramment en RI (TF, TF·IDF, OkapiBM25) basées sur une fréquence binaire et d'utiliser le TF sur la fréquence

$$\begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ D_n \end{array} \begin{array}{c} \begin{array}{cccc} T_1 & T_2 & \dots & T_t \end{array} \\ \begin{pmatrix} w_{11} & w_{21} & \dots & w_{t1} \\ w_{21} & w_{22} & \dots & w_{t2} \\ \vdots & \vdots & & \vdots \\ w_{n1} & w_{n2} & \dots & w_{tn} \end{pmatrix} \end{array}$$

TABLE 5.3 – Modèle vectoriel : matrice document-par-tuiles

proportionnelle (TFp). Concernant la recherche d'information basée sur le tuilage nous proposons d'utiliser l'approche vectorielle de Salton et le produit scalaire. Étant donné que l'information peut être représentée sous plusieurs échelles, un tuilage multi-échelle permet d'utiliser l'index de tuiles le plus adapté au grain de la requête d'un utilisateur.

Pour évaluer l'intérêt de cette approche d'uniformisation, nous avons effectué plusieurs expérimentations pour l'information spatiale et temporelle qui sont présentées dans le chapitre 8. Nous avons donc voulu dans un premier temps vérifier que la perte de précision due au tuilage ne dégrade pas de manière trop importante les résultats. Pour cela, nous avons comparé un SRI spatial standard avec un SRI spatial étendu par notre approche d'uniformisation et de même pour le temporel, en utilisant une approche d'évaluation de SRI classique (voir section 2.4.4). Comme nous allons le voir, l'uniformisation spatiale améliore les résultats par rapport à la référence (13 %), et l'uniformisation temporelle retourne des résultats équivalents à la référence. L'autre principale expérimentation réalisée a eu pour but de déterminer quel tuilage et quelle formule de pondération donnent les meilleurs résultats pour l'information spatiale et pour l'information temporelle. Suite à ces résultats, nous allons donc pouvoir nous intéresser dans le chapitre qui suit à la combinaison de ces facettes géographiques, qui est notre objectif principal.

Chapitre 6

Recherche d'information géographique par combinaison de critères

Sommaire

6.1 Introduction

Pour éviter de biaiser la combinaison des données relatives à différentes facettes de l'information géographique (spatial, temporel et thématique), ou de devoir se limiter à des approches de type filtrage (voir chapitre 3), nous nous sommes intéressés à l'uniformisation de ces données. Dans le chapitre précédent, nous avons ainsi proposé une approche générale d'uniformisation que nous avons appliquée aux informations spatiales et temporelles ; elle se veut comparable à la généralisation par troncature ou lemmatisation de termes dans une approche de RI classique. De plus, la mise en œuvre du modèle vectoriel et de formules éprouvées en RI nous assure une réelle homogénéité des démarches de RI appliquées pour chacune des facettes de l'information géographique. Il faut noter que dans ce chapitre, pour le thématique, nous nous limiterons aux approches

utilisées sur les termes en RI classique comme le font la plupart des SRIG présentés dans la section 2.5. Dans ce chapitre, notre but est de combiner les résultats issus de recherche dans ces index indépendants.

	RI Basique	RI Étendue
1 critère (termes)	approches standards (Google, Terrier, Lucene)	approches personnalisables/contraintes (Google, Terrier, Lucene)
N critères (spatial, temporel, thématique)	approches standards (GéoSem, DIGMAP)	approches personnalisables/contraintes (∅)

TABLE 6.1 – La combinaison de critères de recherche en RI & RIG

Comme nous l'avons présenté dans la section 1.3 et comme le rappelle le tableau 6.1, les moteurs de recherche usuels (tels que Google, Terrier ou Lucene) proposent 2 types de formulation du besoin :

- standard : l'utilisateur fournit un ou plusieurs mots-clés ; le système combine les différents ensembles de résultats sans que l'utilisateur puisse intervenir.
- étendue : l'utilisateur fournit plusieurs mots-clés et il spécifie ceux qui sont obligatoires, à exclure, ou encore il peut leur attribuer des poids.

En ce qui concerne l'information géographique, les systèmes supportant une combinaison permettant d'obtenir une liste ordonnée de résultats utilisent des approches de type linéaire standard (GéoSem [BDEH07], DIGMAP [MMBS09]). Néanmoins ces systèmes n'ont pas évalué l'apport de la RIG via ces combinaisons. En effet, en règle générale, dans la littérature de la RIG, les systèmes sont évalués sur leurs performances (*efficiency*), par exemple sur le temps de calcul ou le stockage, et, non pas sur la qualité de leurs résultats (*effectiveness*). Notre hypothèse est que la combinaison de différentes facettes de l'information géographique améliore la pertinence des résultats. Nous avons donc décidé de tester plusieurs approches de combinaisons linéaires standards existantes pour vérifier l'intérêt de combiner ces différentes facettes mais aussi quantifier cet apport.

Nous proposons ensuite une approche de combinaison linéaire étendue similaire à la RI classique pour la RI multicritère (case ∅ du tableau 6.1). Le but est d'offir à un utilisateur une plus grande maîtrise de la combinaison des différents critères. En effet, nous avons identifié plusieurs scénarios de recherche (que nous allons présenter) pour lesquels une approche de type linéaire standard ne permet pas de réaliser ces différents scénarios. Nous avons donc proposé une approche de combinaison linéaire étendue basée sur les contraintes en nous inspirant de celles existantes pour l'aide à la décision multicritère (voir section 3.3).

Enfin, nous constatons qu'il n'existe pas de cadre d'évaluation de Systèmes de Recherche d'Information Géographique (SRIG) comme nous l'avons indiqué dans la section 2.4.4. Pour pouvoir tester et évaluer ces différentes approches de combinaison nous avons donc mis en place un cadre expérimental pour évaluer un SRIG en se basant sur les approches utilisées en RI classique, notamment dans les campagnes TREC.

Tout au long de ce chapitre nous allons illustrer les différentes approches avec les deux exemples suivants :

1. L'utilisateur prépare une randonnée en montagne et cherche à prendre connaissance des documents évoquant les risques accidentels associés. Les documents retournés doivent être centrés sur « la montagne » et « les accidents ». S'ils relatent de balades/randonnées ce sera effectivement un plus non négligeable. Le tableau 1.2 page 10 illustre cette requête. Le SRI Terrier nous permet d'exprimer cette requête par : `+montagne +accident^7 balade^7` (accident et balade auront un poids supérieur à montagne). Cette requête a la particularité suivante : les documents retournés doivent impérativement parler d'accident et de montagne et éventuellement de balade. Toutefois, les documents traitant de balade bénéficieront d'un coefficient de préférence élevé pour un reclassement vers le début de la liste de résultats. Idem pour ceux liés aux accidents. A contrario, tous les documents doivent traiter de la montagne, mais ce critère n'aura pas une influence importante dans le calcul de score, ce qui s'apparente à du filtrage.
2. L'utilisateur s'intéresse aux montagnes des Pyrénées et cherche des descriptions passées de ces montagnes. Il souhaite néanmoins exclure les documents portant sur la commune de Gavarnie et éviter ceux centrés sur les ascensions. Le tableau 1.3 page 10 illustre cette requête.

Ce chapitre présente donc trois contributions. La première consiste à montrer l'apport de la combinaison de l'information géographique et à la quantifier. La deuxième propose une approche originale et générique (utilisable sur divers critères) de combinaison étendue grâce à des contraintes, que nous appliquons à la RIG. La dernière met en œuvre un cadre expérimental d'évaluation de SRIG.

6.2 Combinaisons linéaires standards

Tout d'abord, nous avons opté pour des approches linéaires standards. En effet, elles permettent, à partir des scores de chaque critère, d'obtenir un score unique sans paramétrisation ou réglage. Elles sont ainsi faciles à mettre en place et interviennent uniquement au niveau de calcul des scores. En contrepartie l'utilisateur ne peut pas personnaliser le mode de combinaison.

Dans la littérature de RI, Fox et al. [FS93] ont proposé plusieurs combinateurs : CombMAX, CombMIN, CombSUM, CombANZ, CombMNZ, complétés par la normalisation proposée par Lee [Lee97]. Cette normalisation permet, comme expliqué dans la section 4.2, de borner les scores de similarité calculés par chaque SRI sur l'intervalle [0;1] (voir équation 6.1).

$$\text{score\_normalisé} = \frac{\text{score\_non\_normalisé} - \text{score\_minimum}}{\text{score\_maximum} - \text{score\_minimum}} \quad \text{[Lee97]} \tag{6.1}$$

Le tableau 6.2 présente les différentes formules proposées par Fox et al. [FS93]. CombMAX retourne le meilleur score parmi les différents critères. Donc il suffit qu'un seul

critère soit satisfait pour avoir un score non nul. Il y a compensation totale entre les critères. CombMIN retourne le moins bon score parmi les différents critères. Tous les critères doivent donc être satisfaits sinon CombMIN = 0. Il n'y aucune compensation entre les critères. CombSUM somme les scores des différents critères. CombANZ divise cette somme par le nombre de critères ayant un score non nul, ce qui revient à faire la moyenne des scores non nuls. Pour finir CombMNZ multiplie CombSUM par le nombre de critères ayant un score non nul. Cette approche permet de donner plus de poids aux documents satisfaisant le plus de critères. 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 à l'égard de ces documents. Quelle que soit la formule utilisée, le but est d'obtenir une seule liste de résultats ordonnés par ordre de pertinence. La figure 6.1 illustre le principe avec la fonction CombMNZ. Comme nous pouvons le voir dans cette figure, il y a trois listes de résultats provenant de chaque SRI monocritère. Pour chaque document un score global (agrégé) est calculé par rapport aux différents scores obtenus normalisés et ce score est pondéré par rapport au nombre de SRI qui le retournent (par exemple le document d_8 n'est présent que dans les résultats de deux SRI donc il sera pondéré par 2).

CombMAX	Score maximum
CombMIN	Score minimum
CombSUM	Somme des scores
CombANZ	CombSUM ÷ nombre de scores non nuls
CombMNZ	CombSUM × nombre de scores non nuls

TABLE 6.2 – Formules de combinaisons proposées par Fox et al. [FS93]

Lee a montré que la formule CombMNZ est celle qui donne les meilleurs résultats [Lee97]. Afin de vérifier que ce combinateur est celui qui donne les meilleurs résultats en RIG, comme en RI classique, nous avons choisi de tester également les autres combinateurs proposés (voir chapitre 8 pour l'expérimentation).

Reprenons nos exemples de l'introduction pour illustrer une approche linéaire standard telle que CombMNZ. Pour la requête du premier exemple (tableau 1.2), nous obtenons le classement de la figure 6.2. La normalisation n'est pas nécessaire étant donné que les trois ensembles de résultats sont obtenus avec le même SRI. Comme nous pouvons le constater, malgré le fait que le document d_3 ne réponde pas du tout au critère « balade » il compense avec les deux autres critères qui ont un score élevé. Néanmoins, si nous souhaitons tenir compte des critères fournis par l'utilisateur dans sa requête, c'est à dire retourner des documents contenant obligatoirement « montagne » et « accident » tout en mettant fortement l'accent sur ce dernier, cette approche ne nous permet pas d'exprimer ce type de contrainte. En effet, seuls d_3, d_5, d_6, d_7 et $d8$ devraient être restitués dans ce cas.

Pour la requête du deuxième exemple (tableau 1.3 page 10), la figure 6.3 illustre la combinaison avec CombMNZ. Les documents résultats sont des extraits de notre corpus (tableau 1.1 page 9). Ici aussi, les contraintes de la requête ne peuvent être exprimées. Seuls d_5 et d_6 devraient être retournés.

d	s
d_4	14,5
d_3	12
d_1	0,5

(a) SRI thématique - critère 1

d	s
d_8	150
d_1	120
d_4	80
d_9	−10

(b) SRI spatial - critère 2

d	s
d_8	1
d_4	0,7
d_9	0,5
d_1	0,5

(c) SRI temporel - critère 3

↘ ↓ ↙

Doc d	Score s calculé avec CombMNZ
d_4	$6,0333 = 3 \times \left(\frac{14,5-0,5}{14,5-0,5} + \frac{80+30}{150+30} + \frac{0,7-0,5}{1-0,5}\right)$
d_8	$4,0000 = 2 \times \left(\frac{150+30}{150+30} + \frac{1-0,5}{1-0,5}\right)$
d_1	$1,6777 = 2 \times \left(\frac{0,5-0,5}{14,5-0,5} + \frac{120+30}{150+30}\right)$
d_3	$0,8214 = 1 \times \left(\frac{12-0,5}{14,5-0,5}\right)$
d_9	$0,2222 = 2 \times \left(\frac{-10+30}{150+30} + \frac{0,5-0,5}{1-0,5}\right)$

(d) SRI géographique résultant de la combinaison des trois résultats

FIGURE 6.1 – Principe de combinaison de résultats de recherche avec CombMNZ.

d	s
d_8	0,9
d_4	0,6
d_5	0,4

(a) SRI thématique - balade

d	s
d_6	0,9
d_3	0,7
d_8	0,7
d_1	0,5
d_5	0,3
d_7	0,3
d_4	0,2

(b) SRI thématique - montagne

d	s
d_3	0,9
d_7	0,8
d_2	0,3
d_5	0,2
d_6	0,1
d_8	0,1

(c) SRI thématique - accident

↘ ↓ ↙

Doc d	Score s calculé avec CombMNZ
d_8	$5,1 = 3 \times (0,9 + 0,7 + 0,1)$
d_3	$3,2 = 2 \times (0,7 + 0,9)$
d_5	$2,7 = 3 \times (0,4 + 0,3 + 0,2)$
d_7	$2,2 = 2 \times (0,3 + 0,8)$
d_6	$2,0 = 2 \times (0,9 + 0,1)$
d_4	$1,2 = 2 \times (0,6 + 0,2)$
d_1	$0,5 = 1 \times (0,5)$
d_2	$0,3 = 1 \times (0,3)$

(d) Combinaison des trois ensembles résultats

FIGURE 6.2 – Résultats de l'exemple 1 (tableau 1.2) avec CombMNZ

d	s
d_1	1,00
d_2	1,00
d_3	1,00
d_4	1,00
d_5	1,00
d_6	1,00

(a) SRI Spatial (TF) - Pyrénées

d	s
d_1	1,00
d_2	0,50
d_4	0,50
d_6	0,01
d_5	0,01

(b) SRI Spatial (TF) - Gavarnie

d	s
d_1	1,00
d_2	1,00
d_3	1,00
d_5	1,00

(c) SRI Temporel (TF) - 1800-1900

d	s
d_6	0,77
d_5	0,67
d_4	0,67

(d) SRI Thematique (Terrier) - montagne

d	s
d_2	1,05
d_3	0,78
d_4	0,72
d_6	0,56

(e) SRI Thematique (Terrier) - ascension

Doc d	Score s calculé avec CombMNZ
d_2	$5,98 = 4 \times \left(\frac{1-1}{1-1} + \frac{0,5-0,01}{1-0,01} + \frac{1-1}{1-1} + \frac{1,05-0,56}{1,05-0,56}\right)$
d_6	$4,00 = 4 \times \left(\frac{1-1}{1-1} + \frac{0,01-0,01}{1-0,01} + \frac{0,77-0,67}{0,77-0,67} + \frac{0,56-0,56}{1,05-0,56}\right)$
d_4	$3,29 = 4 \times \left(\frac{1-1}{1-1} + \frac{0,5-0,01}{1-0,01} + \frac{0,67-0,67}{0,77-0,67} + \frac{0,72-0,56}{1,05-0,56}\right)$
d_1	$3,00 = 3 \times \left(\frac{1-1}{1-1} + \frac{1-0,01}{1-0,01} + \frac{1-1}{1-1}\right)$
d_3	$1,35 = 3 \times \left(\frac{1-1}{1-1} + \frac{1-1}{1-1} + \frac{0,78-0,56}{1,05-0,56}\right)$
d_5	$0,00 = 4 \times \left(\frac{1-1}{1-1} + \frac{0,01-0,01}{1-0,01} + \frac{1-1}{1-1} + \frac{0,67-0,67}{0,77-0,67}\right)$

(f) SRI géographique résultant de la combinaison des trois résultats

FIGURE 6.3 – Résultats de l'exemple 2 (tableau 1.3) avec CombMNZ.

Les approches présentées ici nous ont permis de tester la combinaison des différentes facettes de l'information géographique et de vérifier son apport. En effet, elles sont faciles et rapides à mettre en place : elles interviennent uniquement pour le calcul du score global et il n'y a pas de paramétrisation. Comme nous allons le voir dans le chapitre 8, la combinaison des différentes facettes géographiques avec CombMNZ améliore nettement les résultats : le gain de précision est d'environ 66,0 % que ce soit par rapport au thématique, spatial ou temporel seul. Suite à ces résultats significatifs, nous avons entrepris l'étude de combinaisons étendues. L'objectif est de permettre à un utilisateur de personnaliser la combinaison afin que le classement reflète plus encore son besoin.

6.3 Combinaisons linéaires étendues

Les approches linéaires standards présentées dans la section précédente permettent de combiner différents critères. Néanmoins elles ne sont pas suffisantes si nous souhaitons prendre en charge différents scénarios de recherche. En effet, chaque critère (choix) peut être utilisé différemment selon la situation : il est possible de spécifier des exigences (présence obligatoire/facultative), et des préférences (importance du critère).

Dans un premier temps, nous avons utilisé les approches de combinaison par priorité proposées par Costa Pereira et al. [CPDP09a, CPDP09b] qui diffèrent des approches

linéaires standards en permettant à l'utilisateur de classer les critères par ordre de préférence. Puis nous proposons une approche de combinaison étendue via des contraintes qui s'inspire des approches utilisées pour l'aide à la décision.

6.3.1 Combinaisons étendues par niveaux de priorités

Dans la littérature présentée dans la section 3.3 sur l'agrégation de critères, les travaux de Costa Pereira et al. [CPDP09a] ont proposé deux approches permettant une prise en compte de certaines contraintes. Ces travaux proposent d'utiliser un système de priorité pour combiner plusieurs critères.

La première approche, « *Prioritized Scoring Model* » (modèle de score par ordre de priorité), met en œuvre le principe selon lequel le poids du critère le moins important doit être proportionnel au score obtenu pour les critères plus importants. L'évaluation d'un document traite les critères de recherche dans l'ordre et s'arrête au premier critère non satisfait par le document. Ce modèle permet d'exprimer un niveau d'exigence de base : au moins le premier critère de restriction exprimé doit être satisfait, les autres sont autant de critères facultatifs avec une hiérarchie d'importance qui doit être respectée. Reprenons la requête 1 (tableau 1.2) avec comme ordre de priorité : accident > montagne > balade. Le classement obtenu est :

- d_1 : $1\times0 + 0\times0{,}5 + 0\times0 = 0$
- d_2 : $1\times0{,}3 + 0{,}3\times0 + 0\times0 = 0{,}3$
- d_3 : $1\times0{,}9 + 0{,}9\times0{,}7 + 0{,}63\times0 = 1{,}53$
- d_4 : $1\times0 + 0\times0{,}2 + 0\times0{,}6 = 0$
- d_5 : $1\times0{,}2 + 0{,}2\times0{,}3 + 0{,}06\times0{,}4 = 0{,}284$
- d_6 : $1\times0{,}1 + 0{,}1\times0{,}9 + 0{,}09\times0 = 0{,}19$
- d_7 : $1\times0{,}9 + 0{,}9\times0{,}3 + 0\times0 = 1{,}17$
- d_8 : $1\times0{,}1 + 0{,}1\times0{,}7 + 0{,}07\times0{,}9 = 0{,}233$

Soient les résultats suivants, par ordre de pertinence : d_3, d_7, d_2, d_5, d_8, d_6. Nous pouvons noter que certains documents (d_8, d_5, d_6) se retrouvent en fin de classement alors qu'ils étaient mieux classés avec CombMNZ. Prenons le cas de d_8 : il était classé premier avec CombMNZ et ici il est classé avant-dernier. La raison est que le score de d_8 pour le critère « accident » est très faible (0,1) donc avec l'approche « *Prioritized Scoring Model* » les autres critères sont pondérés avec un faible coefficient (même si leur score est élevé). En effet, la requête indique que le critère « accident » est le plus important et donc que son score doit avoir plus d'impact que les autres. L'approche « *Prioritized Scoring Model* » permet donc de réaliser un classement des documents plus en adéquation avec la requête de l'utilisateur. Cette approche permet donc de rendre un critère obligatoire (le premier), mais uniquement un seul. Dans le cas de l'exemple donné, cette approche a pu écarter les documents d_1 et d_4 (critère « accident » non satisfait) mais n'a pas pu écarter le document d_2 qui ne satisfait pas le deuxième critère obligatoire (« montagne »).

La deuxième approche, « *Prioritized And Model* » (modèle ET par ordre de priorité), part du principe selon lequel tous les critères exprimés sont essentiels. Elle ne prend pas en compte des requêtes dans lesquelles un critère serait facultatif. Chaque critère est pondéré automatiquement par rapport à sa place dans l'ordre de priorité. Le score global est en

fait la similarité minimum des différents critères. Reprenons la requête 1 (tableau 1.2) avec comme ordre de priorité : accident > montagne > balade. Le classement obtenu est :

- d_1 : $\min(0^1, 0{,}5^0, 0^0) = 0$
- d_2 : $\min(0{,}3^1, 0^{0,3}, 0^0) = 0$
- d_3 : $\min(0{,}9^1, 0{,}7^{0,9}, 0^{0{,}63}) = 0$
- d_4 : $\min(0^1, 0{,}2^0, 0{,}6^0) = 0$
- d_5 : $\min(0{,}2^1, 0{,}3^{0,2}, 0{,}4^{0{,}06}) = 0{,}2$
- d_6 : $\min(0{,}1^1, 0{,}9^{0,1}, 0^{0{,}09}) = 0$
- d_7 : $\min(0{,}8^1, 0{,}3^{0,8}, 0^{0{,}24}) = 0$
- d_8 : $\min(0{,}1^1, 0{,}7^{0,1}, 0{,}9^{0{,}07}) = 0{,}1$

Soient les résultats suivants, par ordre de pertinence : d_5, d_8. Cette approche ne permettant pas de spécifier des critères facultatifs, les documents d_3, d_6 et d_7 sont écartés alors qu'ils répondent à la requête (le critère « balade » n'est pas obligatoire).

Ces deux approches permettent, contrairement aux approches linéaires standards, de contraindre la combinaison. Elles restent faciles et rapides à mettre en place, étant donné que l'utilisateur ne spécifie qu'un classement d'importance des critères. Dans le premier cas, seul le premier critère est obligatoire, et dans le deuxième ils le sont tous. Les préférences sont de plus calculées automatiquement (1 pour le premier critère et les suivants dépendent du score du précédent). Néanmoins, pour le cas particulier de la requête 1 (tableau 1.2) présentée ici, ces approches ne permettent pas de les exprimer conformément à la demande de l'utilisateur. En effet, si nous souhaitons avoir plusieurs critères obligatoires (mais pas la totalité) ces approches ne permettent pas de le faire. Les expérimentations réalisées révèlent que CombMNZ donne de meilleurs résultats que ces deux approches (voir chapitre 8) dans notre contexte de RIG. Il faut noter que ces approches ont été proposées pour des recherches personnalisées sur les termes et non pas pour la RIG, donc elles ne sont pas adaptées à des requêtes aussi complexes. Nous proposons donc une approche de combinaison étendue via des contraintes que nous allons détailler maintenant.

6.3.2 Combinaisons étendues par niveaux d'exigences, de préférences et d'opérateurs

Nous proposons donc une nouvelle approche, similaire à celles utilisées en agrégation de critères (voir section 3.3), pour mettre en place une recherche étendue multicritère. Le but est de permettre à l'utilisateur de spécifier pour chaque critère une exigence (présence obligatoire/facultative), et une préférence (importance du critère). De plus, nous proposons la possibilité de choisir l'opérateur utilisé pour chaque critère.

6.3.2.1 Descriptif général

Le système de combinaison par contraintes que nous proposons peut ainsi être modélisé comme un quadruplet (C, O, P, E) :

- les choix C = $(c_1, \ldots, c_n)$: un choix est un critère de recherche ;

- les opérateurs $O = (o_1, \ldots, o_n)$ comparant un document D et un choix C. Divers opérateurs sont envisageables : intersection, égalité, inclusion, et proximité. Les opérateurs sont à la charge des SRI utilisés (voir section 7.4 pour ceux que gère notre SRI) ;
- les exigences $E = (e_1, \ldots, e_n)$: obligatoire, facultatif, exclus ;
- les préférences $P = (p_1, \ldots, p_n)$, *avec* $p_i \in \mathbb{R}$, qui permettent de pondérer chaque critère de recherche en fonction de son importance.

Nous avons identifié cinq scénarios applicables à un critère : l'exclusion (filtrage négatif), la dévalorisation, la valorisation, la cible (focus), et le prérequis ou délimitation (filtrage positif). Le tableau 6.3 illustre ces différents scénarios que nous allons maintenant détailler. Dans un premier temps considérons une préférence comme ternaire pour prendre en compte la dévalorisation : 0, 1 ou -1. Nous les noterons dans nos tableaux respectivement N (pour neutre), $+$ et $-$. Pour les exigences nous utiliserons aussi cette syntaxe : N pour facultatif, $+$ pour obligatoire et $-$ pour exclusion.

A - Exclusion (filtrage négatif) Un utilisateur peut souhaiter exclure un critère. Le critère ne doit donc pas être satisfait et ainsi il n'est pas nécessaire de lui attribuer une préférence.

B - Dévalorisation Un utilisateur peut décider de défavoriser un critère pour ne pas exclure les documents contenant ce critère mais plutôt réduire leur score global.

C - Valorisation Un utilisateur peut décider d'inclure un critère facultatif. Si ce critère est présent le document est plus intéressant, sinon c'est sans conséquence. Ainsi un poids est associé à ce critère pour qu'il influence le calcul du score global.

D - Cible (focus) Un critère (ou plusieurs) peut être défini comme le but de la recherche. Il doit être présent et il va intervenir dans le calcul du score global. Si tous les critères sont cibles avec un poids équivalent cela revient à utiliser une moyenne arithmétique.

E - Prérequis - Délimitation (filtrage positif) Un utilisateur peut décider de se servir d'un critère comme délimitateur et ainsi que ce critère n'intervienne pas dans le calcul du score global. Le critère doit être présent (obligatoire) mais sans degré de préférence.

Il faut noter que certaines combinaisons sont incohérentes (par exemple exigence $-$ et préférence $+$). Il sera donc nécessaire de réaliser des contrôles.

Ainsi nous voyons l'intérêt d'utiliser cette approche par contraintes pour que l'utilisateur puisse paramétrer la combinaison selon son besoin. Concernant le calcul du score global, nous proposons d'utiliser une somme pondérée (voir équation 6.2). Cette somme pondérée permet de prendre en compte les différentes préférences fournies afin que chaque critère influence différemment (si besoin) le score global. Concernant les exigences, si un

Critères	Exigences			Préférences			
	−	N	+	−	N	+	
A	✓				✓		Exclusion (Filtrage négatif)
B		✓		✓			Dévalorisation
C		✓				✓	Valorisation
D			✓			✓	Cible (Focus)
E			✓		✓		Prérequis (Filtrage positif)

TABLE 6.3 – Scénarios de recherche possible

critère obligatoire n'est pas présent ou au contraire un critère à exclure est présent, il n'est pas nécessaire de calculer un score car le document ne doit pas être retourné.

$$\text{score}(d_i) = \begin{cases} 0 \ si \ exigences\_non\_satisfaites(d_i, E) \\ \frac{1}{\Sigma_{i=1}^{n} p_i > 0} * \Sigma_{i=1}^{n}(p_i * o_i(d_i, c_i)) \ sinon \end{cases} \tag{6.2}$$

6.3.2.2 Exemples d'application

L'approche de combinaison par contraintes que nous proposons peut s'appliquer à différents critères préalablement homogénéisés (par exemple, via la démarche proposée dans le chapitre précédent). Étant donné que nous souhaitons combiner les données issues des différentes facettes de l'information géographique, nous allons illustrer cette approche via les exemples donnés dans l'introduction de ce chapitre.

Choix	balade	montagne	accident
Opérateur	égalité	égalité	égalité
Exigence	N	+	+
Préférences	+	N	+

TABLE 6.4 – Requête 1 : Choix, Opérateurs, Préférences et Exigences

La requête 1 (tableau 1.2) peut s'écrire sous la forme du tableau 6.4. Ici nous avons trois critères : (a) « balade » (valorisation), (b) « montagne » (prérequis), (c) « accident » (cible). Nous obtenons ce classement :

- $d_8 : \frac{1}{1+0+1} \times (1 \times 0,9 + 0 \times 0,7 + 1 \times 0,1) = 0,50$
- $d_3 : \frac{1}{1+0+1} \times (1 \times 0 + 0 \times 0,7 + 1 \times 0,9) = 0,45$
- $d_7 : \frac{1}{1+0+1} \times (1 \times 0 + 0 \times 0,3 + 1 \times 0,8) = 0,40$
- $d_5 : \frac{1}{1+0+1} \times (1 \times 0,4 + 0 \times 0,3 + 1 \times 0,2) = 0,30$
- $d_6 : \frac{1}{1+0+1} \times (1 \times 0 + 0 \times 0,9 + 1 \times 0,1) = 0,05$

Les documents d_1, et d_4 sont écartés car ils ne respectent pas l'exigence du critère (c) (« accident ») et le document d_2 est également écarté car il ne respecte pas l'exigence du critère (b) (« montagne »). Contrairement aux approches présentées auparavant (linéaires standards ou par priorités), nous obtenons tous les documents répondant aux contraintes de la requête et uniquement ceux là. Néanmoins, concernant le classement, nous remarquons que d_8 est en tête alors qu'il répond très bien au critère optionnel (a)

mais très peu à la cible (c). Cet effet de bord est dû au fait que les préférences sont exprimées de manière ternaire (−1, 0 ou 1). Ce paramétrage ne permet pas d'exprimer assez finement les contraintes. Nous allons voir par la suite comment supprimer cet effet de bord.

De même, nous pouvons exprimer la requête 2 (tableau 1.3) portant sur la RIG avec l'approche par contraintes : tableau 6.5.

Contraintes	Spatial	Spatial	Temporel	Thème	Thème
C	Pyrénées	Gavarnie	1800-1900	montagnes	ascension
O	intersection	égalité	intersection	égalité	égalité
E	+	-	N	+	N
P	N	N	+	+	-

TABLE 6.5 – Requête 2 : Choix, Opérateurs, Préférences et Exigences

Ici nous avons les cinq cas de figure : (a) « Pyrénées » (prérequis), (b) « Gavarnie » (exclusion), (c) « 1800-1900 » (valorisation), (d) « montagnes » (cible), (e) « ascension » (dévalorisation). Nous obtenons ce classement :

- $d_5 : \frac{1}{0+1+1} \times (0 \times 1 + 1 \times 1 + 1 \times 0,67) = 0,83$
- $d_6 : \frac{1}{0+1+1} \times (0 \times 1 + 1 \times 0 + 1 \times 0,77 - 1 \times 0,56) = 0,10$

Seuls deux documents répondent aux différents critères : d_5 et d_6. Les documents d_1, d_2 et d_3 sont écartés car ils ne satisfont pas l'exigence du critère (d) (« montagne »), d_4 est également écarté car il ne satisfait pas l'exigence du critère (b) (« Gavarnie »). Ici d_6 est classé loin derrière d_5 car il répond au critère (e) (« ascension ») et il est donc pénalisé.

Concernant les préférences, nous avons présenté une approche ternaire (+, N, −). Comme nous avons pu le voir, cette approche n'est pas suffisamment satisfaisante. Nous proposons alors une nouvelle version de la combinaison par contraintes intégrant une approche proportionnelle : préférence entre 0 et 1 pour chaque critère. Reprenons le dernier exemple. En choisissant des préférences proportionnelle, nous pouvons exprimer la requête sous la forme du tableau 6.6.

Contraintes	Spatial	Spatial	Temporel	Thème	Thème
C	Pyrénées	Gavarnie	1790-1860	montagnes	escalade
O	intersection	égalité	intersection	égalité	égalité
E	+	-	N	+	N
P	0	0	0,8	0,7	-0,5

TABLE 6.6 – Requête 2 : Choix, Opérateurs, Préférences proportionnelles et Exigences

Nous obtenons ce classement :

- $d_5 : \frac{1}{0+0,8+0,7} \times (0 \times 1 + 0,8 \times 1 + 0,7 \times 0,67) = 0,85$
- $d_6 : \frac{1}{0+0,8+0,7} \times (0 \times 1 + 0,8 \times 0 + 0,7 \times 0,77 - 0,5 \times 0,56) = 0,17$

Ici nous ne constatons pas de changements dans le classement. En effet nous avons peu de documents résultats et le classement s'avère en adéquation avec la requête de l'utilisateur. Par contre, si nous reprenons l'exemple de la requête 1 (tableau 1.2). En

choisissant des préférences proportionnelles, nous pouvons exprimer la requête sous la forme du tableau 6.7. Nous obtenons alors ce classement :

- $d_3 : \frac{1}{0,1+0+0,9} \times (0,1 \times 0 + 0 \times 0,7 + 0,9 \times 0,9) = 0,81$
- $d_7 : \frac{1}{0,1+0+0,9} \times (0,1 \times 0 + 0 \times 0,3 + 0,9 \times 0,8) = 0,72$
- $d_5 : \frac{1}{0,1+0+0,9} \times (0,1 \times 0,4 + 0 \times 0,3 + 0,9 \times 0,2) = 0,22$
- $d_8 : \frac{1}{0,1+0+0,9} \times (0,1 \times 0,9 + 0 \times 0,7 + 0,9 \times 0,1) = 0,18$
- $d_6 : \frac{1}{0,1+0+0,9} \times (0,1 \times 0 + 0 \times 0,9 + 0,9 \times 0,1) = 0,09$

Choix	balade	montagne	accident
Opérateur	égalité	égalité	égalité
Exigence	N	+	+
Préférences	0,1	0	0,9

TABLE 6.7 – Requête 1 : Choix, Opérateurs, Préférences proportionnelles et Exigences

Ici nous constatons une modification du classement : d_8 auparavant premier, se retrouve avant-dernier. En effet, il répond très bien au critère facultatif (a) (« balade ») mais très peu à la cible (c) (« accident »). En utilisant une pondération proportionnelle, plus en adéquation avec cette requête (0,1 pour le critère (a) et 0,9 pour le critère (c) au lieu de 1 pour chaque), le score du document d_8 devient 0,18 au lieu de 0,5 auparavant. De même, d_3 qui répond très bien au critère (c) (0,9) voit son score passer de 0,45 à 0,81. L'utilisation de préférences proportionnelles permet donc d'améliorer la puissance d'expression de la requête et, par conséquent, le classement des résultats.

La principale limite de cette approche par contrainte est qu'il n'est pas possible de réaliser des opérations sur des groupes de critères. Par exemple il n'est pas possible d'exprimer « montagner ET (Alpes OU Pyrénées) ».

Dans cette section, nous avons proposé une approche originale de combinaison par contraintes permettant à l'utilisateur de personnaliser la combinaison des différents critères. La dernière section présente notre proposition de cadre d'évaluation d'un SRI géographique afin de pouvoir tester et comparer ces différentes approches de combinaison.

6.4 Cadre expérimental d'évaluation d'un SRI Géographique

Comme nous l'avons signalé dans la section 2.4.4, il n'existe pas de cadre d'évaluation de tel SRI multifacettes. La plupart se limitent à une facette de l'information géographique ou tout au plus à deux (spatial et thématique). Néanmoins pour évaluer et comparer nos approches de combinaison nous avons besoin d'un tel cadre. C'est la raison pour laquelle nous proposons ici un cadre expérimental pour évaluer un SRI Géographique (ce cadre a été publié dans les conférences INFORSID'10 [PCSH10a] et ECDL'10 [PCSH10b]). 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 6.4.1 détaille la constitution d'une collection de test couvrant les trois facettes, puis la section 6.4.2 expose l'analyse des résultats de SRI permettant de comparer leur efficacité.

6.4.1 Constitution d'une collection de test pour évaluer la recherche d'information géographique

Dans la littérature, notamment dans TREC [Har05], 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. [BV00] 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 [VH05].
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* [SJvR75]. 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 facettes 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 facettes. Par exemple, un topic pourrait avoir pour titre *« Transhumance dans les Alpes au* XIX^{e} *siècle »* et pour narration *« 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 facettes : l'aspect thématique classiquement considéré est complété par des éléments spatiaux et temporels ;
3. des *qrels* par facette où l'assesseur évalue l'adéquation entre chacune des trois facettes considérées (thématique, spatiale et temporelle) et le document. Notons que la seule présence des trois facettes 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 facettes. Cette adéquation est actuellement booléenne pour ne pas surcharger les assesseurs ; ce choix rejoint les observations de Bucher et al. [BCJ+05] qui soulignent que les jugements graduels par critère 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 facettes pertinentes et d'autre part la pertinence globale. Notons qu'aucune hypothèse n'est faite sur l'importance relative des facettes, 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.

6.4.2 Protocole d'analyse comparative de 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 monofacette : thématique (Th), spatial (S) et temporel (Te) ;
- SRI bifacettes : Th+S, Th+Te et S+Te permettant de mesurer l'apport de chacune des facettes dans l'efficacité du SRI ;
- SRI géographique combinant les trois facettes : 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 (voir section 2.4.4). Ces dernières requièrent des *qrels* booléens [MRS08, ch. 8]. Or, dans le protocole expérimental proposé, les *qrels* sont graduels afin de juger plus finement la pertinence d'un document par rapport aux trois facettes 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. [JK02] et utilisée notamment dans le cadre de la campagne d'évaluation TREC-9 pour la tâche Web caractérisée par des *qrels* graduels [Voo01]. Cette mesure implémente deux principes. D'une part, les documents très pertinents ($v \to 4$ dans notre cas) sont plus intéressants que les documents peu pertinents ($v \to 1$). D'autre part, un document a d'autant moins d'intérêt pour l'utilisateur

s'il est loin 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 $MANDCG$ des n valeurs de $NDCG$, fournissant ainsi la mesure globale de performance du SRI.

Au niveau global, les différences observées $\langle m_i^1 - m_j^1, \ldots, m_i^n - m_j^n \rangle$ 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 $\forall t \in [1;n]\ m_i^t \not\geqslant m_j^t$). Bien que nécessitant théoriquement une distribution normale des données, Hull [Hul93] précise que ce test est en pratique robuste aux violations de cette condition. Par ailleurs, Sanderson et al. [SZ05] 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 [Hul93]. Plus la valeur p est petite, plus la différence est significative.

Nous avons donc présenté dans cette section un cadre expérimental pour évaluer un SRI géographique. La mise en œuvre de ce cadre et les expérimentations sont présentées dans le chapitre 8.

6.5 Conclusion

Dans ce chapitre nous avons présenté plusieurs contributions. La première consistait à vérifier l'hypothèse que la combinaison des résultats issus des SRI dédiés aux différentes facettes de l'information géographique améliore la pertinence des résultats. Pour vérifier cette hypothèse nous avons utilisé des combinateurs linéaires. L'intérêt majeur des approches de type linéaire standards est qu'elles permettent de calculer directement un score global à partir des scores de chaque critère (maximum, minimum, moyenne, ...). Néanmoins, l'utilisateur ne peut aucunement paramétrer la combinaison.

Nous avons ensuite proposé une approche originale et générique de combinaison étendue par contraintes, qui s'inspire de celles utilisées pour l'aide à la décision et que nous avons appliqué à la RIG. Cette approche propose à l'utilisateur de spécifier la combinaison. Il peut ainsi choisir si un critère est obligatoire, optionnel ou à exclure, l'importance de chaque critère, et l'opérateur à appliquer. Ces approches par contraintes permettent ainsi d'envisager divers scénarios de recherche pour chaque critère : délimitation (filtrage positif), exclusion (filtrage négatif), cible, valorisation et dévalorisation.

Enfin, nous avons proposé un cadre d'évaluation de SRI géographique pour pouvoir évaluer et comparer différentes approches de combinaisons. Nous avons mis en place ce cadre avec notre prototype PIV et le SRI Terrier comme nous le verrons dans le cha-

pitre 8. Avant de détailler ces expérimentations, le chapitre suivant présente les différents prototypes mis au point et utilisés dans le cadre de ces travaux.

Chapitre 7

Implémentations

Sommaire

7.1 Introduction

Dans les deux chapitres précédents, nous avons présenté les approches d'uniformisation et de combinaison que nous proposons. Pour tester ces approches, nous avons mis en place plusieurs prototypes. PIV^2 se base sur les index de précision maximale générés par PIV, uniformise les données spatiales et temporelles qui y sont contenues et crée de nouveaux index. PIVcomb, quant à lui, permet de réaliser des recherches géographiques sur plusieurs critères et de personnaliser la manière dont la combinaison est réalisée.

Pour évaluer et comparer différentes approches de combinaison, nous avons aussi mis au point un prototype d'évaluation d'un RIG, PIVasse, que nous avons utilisé afin de permettre à des assesseurs de juger les résultats issus de notre corpus sur les trois facettes de l'information géographique (spatial, temporel et thématique). Pour nous aider à sélectionner les requêtes en vue de l'évaluation, un dernier prototype a été mis au point, PIVone. Il permet de construire une requête et de vérifier que chaque SRI renvoie des résultats. Il permet, ensuite, de générer le *pool* de documents résultats associé.

Dans ce chapitre nous allons donc présenter ces différents prototypes.

7.2 PIV : Système de Recherche d'Information Géographique dans des documents textuels

Avant de présenter notre système PIV[2], il est nécessaire d'introduire le système sur lequel il s'appuie : PIV. PIV est constitué de deux chaînes de traitement : une pour le spatial et une pour le temporel. Chaque chaîne de traitement génère un index spécialisé. Les chaînes spatiale et temporelle 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 (ESA) tandis que « au nord du gave de Pau » est annoté ES relative (ESR) – relation spatiale d'orientation [GSE+08] ; de même « le printemps 1840 » est annoté ET absolue (ETA) tandis que « vers le printemps 1840 » est annoté ET relative (ETR) – relation temporelle d'adjacence [LGS07]. Ainsi, la création des index spatial et temporel est réalisée en deux étapes. La première étape consiste à extraire les ES et les ET à l'aide d'une chaîne de traitement syntaxico-sémantique [GSE+08]. Cette chaîne est supportée par la plateforme LinguaStream [BCEM03, WB05]. 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 [LGS07] et des géométries aux ES [GSE+08], à l'aide des opérateurs spatiaux du SIG PostGIS [33].

Chaque index est constitué de deux tables (stockées une base de données PostgreSQL): une contenant les informations extraites des documents (voir tableau 7.1) et une contenant toutes les représentations distinctes rencontrées (voir tableau 7.2).

champs	intitulé
doc_id	identifiant du document
par_id	identifiant du paragraphe
ent_id	identifiant de l'entité
label	syntagme correspondant à l'entité
type	type de l'entité (absolue ou relative)
category	catégorie de l'entité (commune, mois, ...)
rep_id	identifiant de la représentation

TABLE 7.1 – Table de l'index contenant les informations extraites

champs	intitulé
rep_id	identifiant de la représentation
rep	représentation

TABLE 7.2 – Table de l'index contenant les représentations

33. http://postgis.refractions.net

7.3 PIV² (« PIVsquare ») : uniformisation des critères

Le prototype PIV² est une extension du prototype PIV. Il se base sur les index générés par PIV. PIV² utilise aussi PostgreSQL et PostGIS. L'application est développée en Java. Le paramétrage de l'application se fait via un fichier XML. Comme l'illustre le listing 7.1, il est possible de spécifier les différents répertoires dans lesquels sont stockés les résultats, les processus à activer (spatial et/ou temporel), les accès aux bases de données, ...

Listing 7.1 – Exemples de paramètres configurables de PIV²

```
<properties>

<!-- General Configuration -->
   <entry key="RESULT">results/</entry>   <!-- Ou vont les resultats -->
   <entry key="RESULT_SPATIAL">spatial/</entry>
   <entry key="RESULT_TEMPORAL">temporal/</entry>

<!-- Criteria -->
   <entry key="SPATIAL">true</entry>   <!-- Activer/Desactiver spatial -->
   <entry key="TEMPORAL">true</entry>  <!-- Activer/Desactiver temporel -->

<!-- Logs -->
   <entry key="LOGS">true</entry>
   <entry key="LOGS_FILE">logs.txt</entry>
   <entry key="LOGS_ERRORS">logs_error.txt</entry>

<!-- Postgres -->
   <entry key="PGSQL_SERVER">localhost</entry>
   <entry key="PGSQL_PORT">5432</entry>
   <entry key="PGSQL_DB">PIVs</entry>
   <entry key="PGSQL_LOGIN">login</entry>
   <entry key="PGSQL_PASS">mdp</entry>

...
</properties>
```

Listing 7.2 – Exemples de spécification de tuilage

```
 <properties>

   <!-- city segmentation -->
   <entry key="GRID_SPATIAL">city</entry>
   <entry key="GRID_SPATIAL_SIZE"></entry>

   <!-- 100x100 grid segmentation -->
   <entry key="GRID_SPATIAL">grid</entry>
   <entry key="GRID_SPATIAL_SIZE">100</entry>


  <!-- month segmentation -->
  <entry key="GRID_TEMPORAL">month</entry>
  <entry key="GRID_TEMPORAL_SIZE"></entry>

  <!-- segmentation with tiles of 10 days -->
  <entry key="GRID_TEMPORAL">grid</entry>
  <entry key="GRID_TEMPORAL_SIZE">10</entry>

</properties>
```

Concernant l'indexation, le prototype permet de générer les différents tuilages souhaités : existants (par exemple tuilage administratif départemental ou calendaire hebdomadaire), ou réguliers. Il faut alors spécifier le maillage souhaité (100x100, 1000x1000, de 10 jours, 40 jours, ...). Le listing 7.2 illustre la spécification de différents tuilages : tuilage communal, tuilage 100x100, tuilage mensuel et tuilage de 10 jours. PIV^2 génère ainsi de nouveaux index, spatiaux ou temporels, stockés dans PostgreSQL.

Chaque index est constitué de deux tables : une contenant le tuilage (voir tableau 7.3) et une contenant les liaisons tuiles-documents avec les poids associés (voir tableau 7.4). Les différents calculs de pondération sont stockés dans la seconde table (tableau 7.4) et associés au triplet (tuile, document, paragraphe), étant donné que l'unité documentaire est le paragraphe. Seul l'IDF est stocké dans la première table (tableau 7.3) car il est lié uniquement à une tuile pour la totalité du corpus. L'ajout ou la suppression d'un document implique donc d'insérer ou de retirer des entrées dans la seconde table (tableau 7.4) et de mettre à jour l'IDF des tuiles impactées dans la première (tableau 7.3). Donc comme nous venons de le voir, pour les calculs de poids des tuiles, quatre formules ont été mises en place (explicitées dans la section 5.3) : TF, TF.IDF, OkapiBM25 et TFp. Il faut noter que pour chaque triplet (tuile, document, paragraphe), la liste des identifiants des représentations (spatiales ou temporelles) impactant la tuile est stockée. Cette liste de représentations va permettre d'enrichir les résultats mais aussi de proposer plusieurs types d'opérateurs de recherche.

champs	intitulé
tile_id	identifiant de la tuile
geom / period	géométrie de la tuile (spatial) / intervalle de la tuile (temporel)
idf	IDF de la tuile par rapport au corpus

TABLE 7.3 – Table de l'index contenant le tuilage

champs	intitulé
tile_id	identifiant de la tuile
doc_id	identifiant du document
par_id	identifiant du paragraphe
idrep_list	liste des identifiants des représentations ayant impacté la tuile
tf	TF associé au triplet (tuile, document, paragraphe)
tfidf	TF.IDF associé au triplet (tuile, document, paragraphe)
okapi	Okapi associé au triplet (tuile, document, paragraphe)
tfp	TFp associé au triplet (tuile, document, paragraphe)

TABLE 7.4 – Table de l'index contenant les liaisons tuiles-documents et les poids associés

Concernant l'interrogation, PIV^2 récupère la(les) représentation(s) spatiale et/ou temporelle associée(s) à la requête et, ensuite, chaque représentation est projetée sur des tuiles comme le montre la figure 7.1. Pour chaque tuile, le système récupère un ensemble de documents et les poids qui leurs sont associés. Nous utilisons une approche

vectorielle [Sal71] et le score d'un document est calculé via une approche de type produit scalaire.

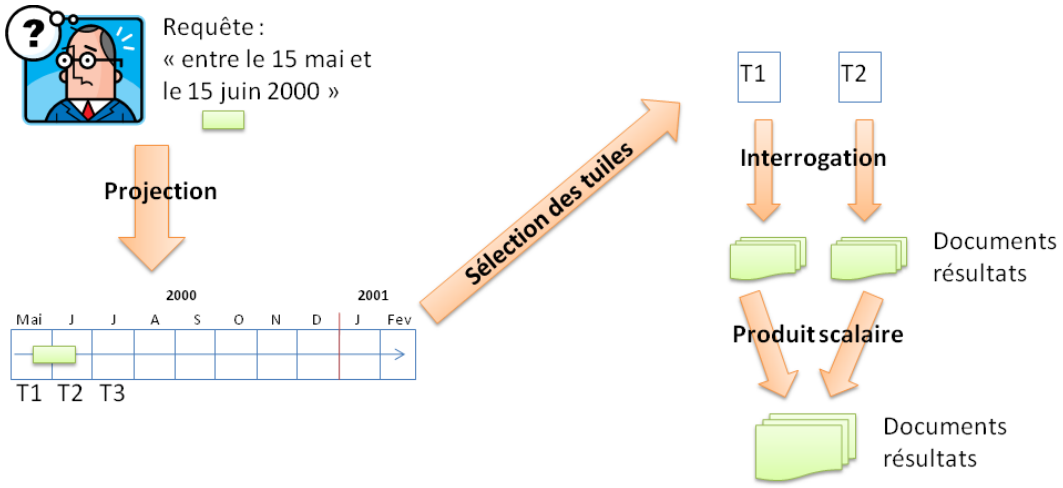


FIGURE 7.1 – PIV² : interrogation par intersection

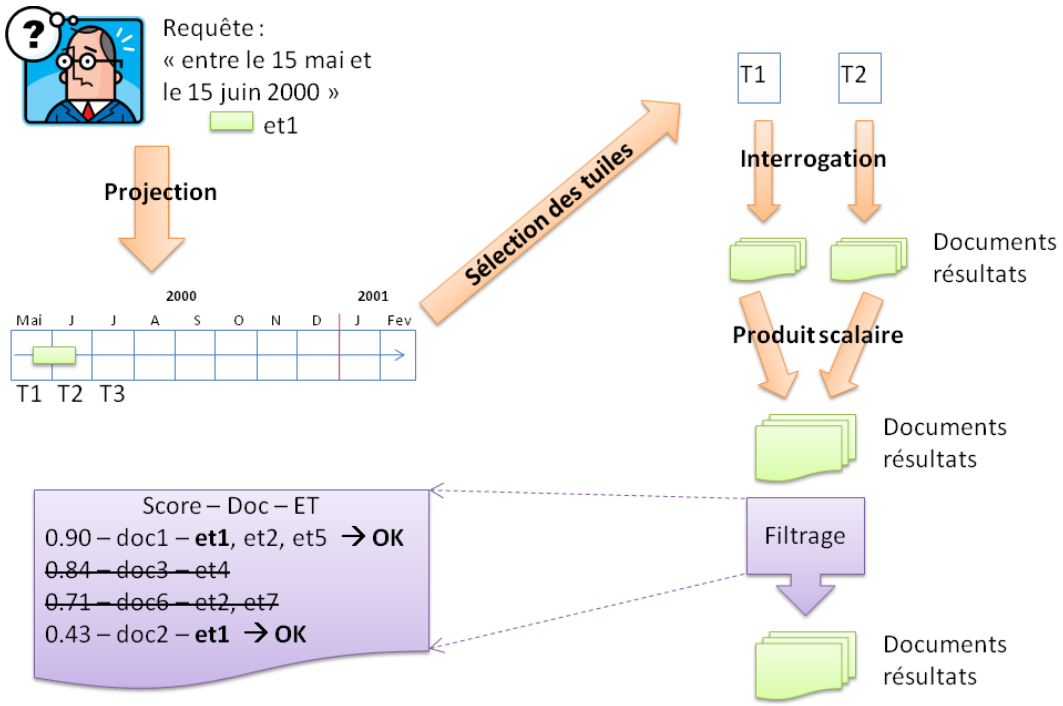


FIGURE 7.2 – PIV² : interrogation par égalité

Parmi les opérateurs proposés dans notre approche de combinaison, PIV^2 supporte l'intersection et l'égalité :

- égalité : information restituée doit être identique à celle demandée
- intersection : information restituée peut englober, être inclue dans, identique à ou recouvrir partiellement celle demandée.

L'intersection consiste à récupérer les tuiles qui intersectent la représentation de la requête de l'utilisateur, comme illustré sur la figure 7.1. L'égalité est plus difficile à gérer car le système est censé être transparent pour l'utilisateur, il ne sait pas quel tuilage est utilisé ni comment. Or le but de l'uniformisation est de convertir les différentes informations en un seul type : la tuile. Donc si un utilisateur souhaite utiliser l'opérateur égalité pour une information spatiale ou temporelle, théoriquement une approche de tuilage ne permet pas de le faire. Néanmoins comme le montre la figure 7.2 il est possible de filtrer les résultats de l'intersection pour ne conserver que ceux contenant la représentation égale à celle demandée, en utilisant les identifiants des représentations associées à chaque tuile. A titre de comparaison nous avons testé une requête via les deux opérateurs : « Pyrénées au XX^e siècle ». Comme nous pouvons le voir dans le tableau 7.5, l'égalité réduit considérablement le nombre de résultats.

Opérateur	Spatial	Temporel
Intersection	2580	211
Egalité	426	1

TABLE 7.5 – Comparaison du nombre de résultats obtenus pour chaque opérateur avec une requête donnée

PIV^2 a été mis en place pour uniformiser les informations spatiales et temporelles. Il se base sur les index générés par PIV, mais pourrait très bien s'appliquer sur des index spatiaux ou temporels générés par d'autres SRIG. PIV^2 produit de nouveaux index spatiaux et temporels sur lesquels nous allons nous baser pour réaliser la combinaison.

7.4 PIVcomb : combinaison par contraintes

PIVcomb permet de soumettre une requête comprenant plusieurs critères géographiques et de spécifier des contraintes liées à chacun de ces critères. Pour l'instant, cette spécification se fait via un fichier XML. Pour chaque critère, l'utilisateur indique différents paramètres :

- l'exigence : critère obligatoire (1), optionnel (0) ou à exclure (-1) ;
- la préférence : valeur comprise entre -1 et 1 ;
- l'opérateur : intersection, égalité, inclusion, proximité (sous réserve que le SRI le gère) ;
- le texte à soumettre au SRI ;
- le SRI à utiliser.

Le listing 7.3 illustre un exemple de requête que prend PIVcomb en entrée. Ce système appelle ainsi le SRI indiqué pour chaque critère et récupère la liste des documents

résultats. À partir de ces différentes listes, PIVcomb génère la liste des documents répondant à la requête complète, ordonnés par score de pertinence.

Listing 7.3 – Exemple de requête contrainte

```
<requete>
 <critere>
  <texte>les Pyrenees</texte>
  <type>spatial</type>
  <exigence>1</exigence>
  <preference>0.8</preference>
  <operateur>intersection</operateur>
 </critere>
 <critere>
  <texte>Gavarnie</texte>
  <type>Spatial</type>
  <exigence>-1</exigence>
  <preference>0</preference>
  <operateur>egalite</operateur>
 </critere>
 <critere>
  <texte>entre 1800 et 1900</texte>
  <type>temporel</type>
  <exigence>0</exigence>
  <preference>0.7</preference>
  <operateur>intersection</operateur>
 </critere>
```

Concernant les résultats obtenus, ils peuvent être riches. Outre le score global du document, il est possible de récupérer les scores relatifs à chaque critère ainsi que les tuiles concernées (voire les entités spatiales, temporelles ou termes pertinents dans le document). L'utilisateur peut ainsi connaitre les éléments qui ont influencé le score global d'un document.

7.5 Outils pour expérimentations

Afin d'évaluer et de comparer différents scénarios de combinaisons, nous avons conçu, développé et mis au point plusieurs prototypes : PIVone pour sélectionner les requêtes et leur *pool* de documents, et, PIVasse pour faciliter le travail des assesseurs.

7.5.1 PIVone (« pivoine ») : vérification et sélection des requêtes

Comme nous l'avons vu dans la section 2.4.4, il est indispensable d'évaluer un système afin de connaître ses performances et l'améliorer. Il faut donc un corpus de référence, des topics ou requêtes et les jugements de pertinences associés à ces requêtes (qrels). Une fois le corpus sélectionné, il faut choisir un nombre minimum de 25 requêtes. Étant donné que l'objectif est d'évaluer un SRIG, ces requêtes doivent porter sur plusieurs facettes de l'information géographique.

Pour vérifier que la requête restitue des résultats sur chacune des facettes ainsi que sur la combinaison, nous avons créé une interface web permettant de soumettre une requête puis de voir le nombre de résultats obtenus. Comme l'illustre la figure 7.3 (capture écran

de PIVone), le système renvoie le nombre de résultats pour chaque facette ainsi que pour chaque paire (spatial+temporel, temporel+thématique, spatial+thématique) et pour la totalité des critères. Cette interface peut utiliser les chaînes spatiale et temporelle de PIV et PIV^2 (au choix) et pour le thématique le SRI Terrier. Il est possible de soumettre une requête sur une, deux ou les trois facettes de l'information géographique. Le système croise ensuite les différentes listes de résultats pour ne conserver que ceux présents dans chacune des listes (intersection). Il n'y a pas de calcul de score ou de classement de résultats. L'objectif est uniquement de vérifier que la requête produit des résultats pour chaque facette ainsi que pour l'ensemble de la requête. Nous pouvons par conséquent retirer les requêtes trop ambigües, c'est à dire retournant trop peu de résultats [San10].

Cette interface Web a été réalisée en (X)HTML, CSS et PHP associés à une base de données MySQL. Les applications PIV et PIV^2 sont utilisées via des archives JAR pouvant ainsi être appelées via une page Web. Le système permet de récupérer pour chaque requête le *pool* de résultats, c'est à dire la liste des documents résultats des différents SRI (les doublons ayant été retirés), tel qu'expliqué dans la section 6.4.1.

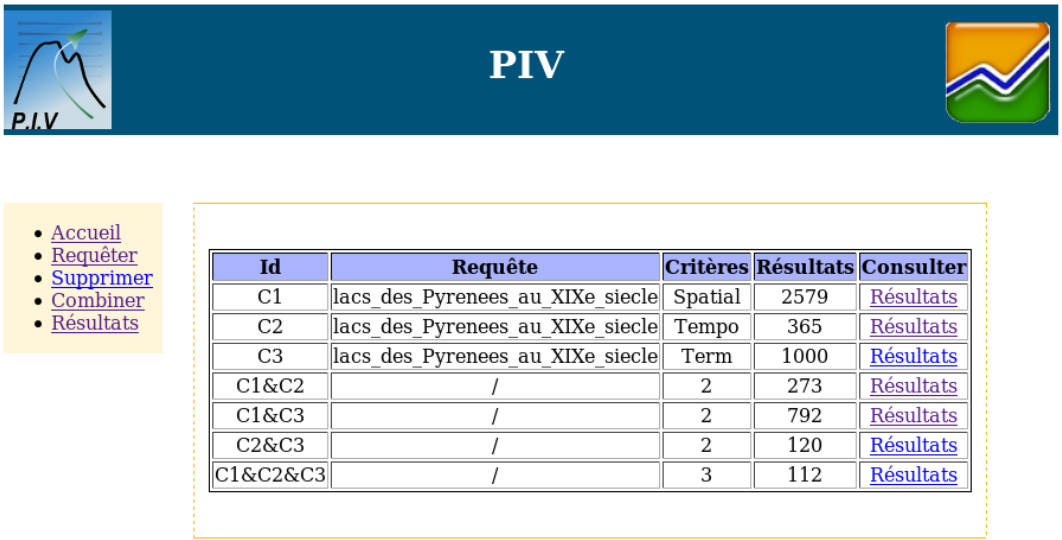


FIGURE 7.3 – PIVone : résultats d'une requête

Cette interface nous a donc permis de voir si, pour une requête donnée, le système renvoie des résultats pour chaque facette et de récupérer le *pool* de documents associé. Une bonne requête ne doit pas être ambigü et retourner un minimum de résultats (si trop peu de résultats, elle est trop difficile). Pour vérifier que les résultats sont effectivement pertinents, il est nécessaire de passer par une phase de jugements manuels (*assessment*).

7.5.2 PIVasse : Évaluations/*Assessment*

Comme expliqué dans la section 6.4, il n'existe pas de cadre d'évaluation pour des SRI géographiques. Pour évaluer notre approche de combinaison géographique nous avons donc proposé un cadre et l'avons mis en œuvre pour tester différentes approches de combinaison [PCSH10a, PCSH10b]. Une fois les requêtes sélectionnées et les listes de résultats obtenues il est nécessaire de les comparer avec des jugements manuels afin d'évaluer le système. Nous avons donc dû mettre en place une plateforme pour permettre à des personnes (« les assesseurs ») de réaliser leurs jugements sur chaque facette et sur la globalité (la présence des trois facettes n'impliquant pas qu'elles soient liées) (voir section 6.4.1 page 95).

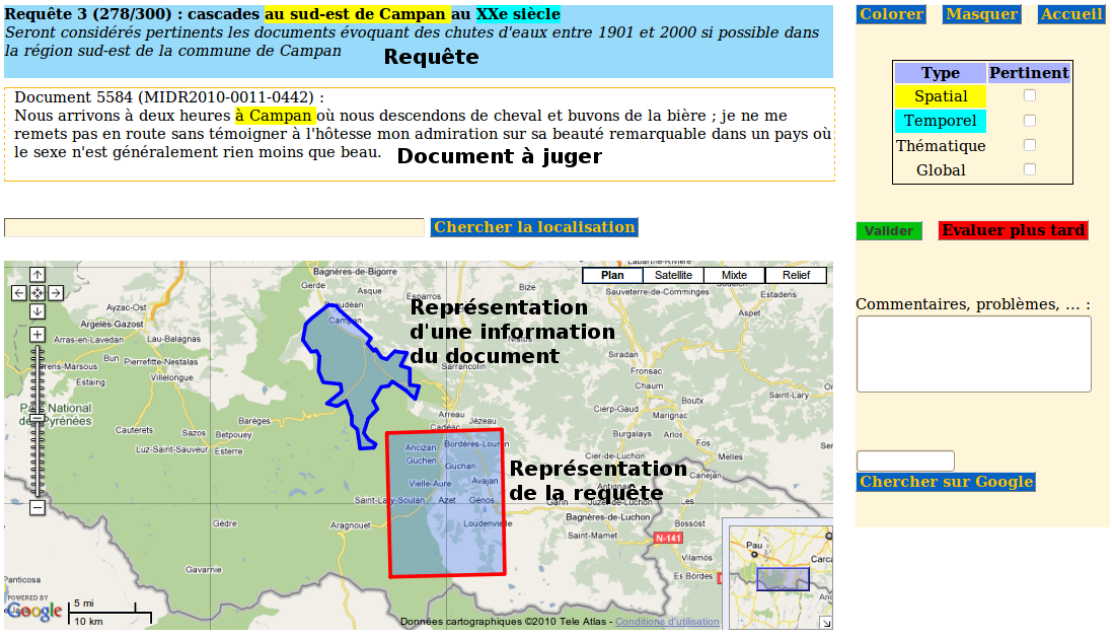


FIGURE 7.4 – PIVasse : évaluation d'un document

L'objectif de la plateforme PIVasse est de permettre aux assesseurs de juger le *pool* de documents associé à chaque requête mais aussi de faciliter ce travail. La capture de la figure 7.4 illustre l'interface de PIVasse. La requête apparaît en haut à gauche, et le document à juger juste en dessous. Les informations spatiales et temporelles y sont surlignées. Cette annotation étant réalisée de manière automatique il subsiste certaines erreurs (des termes non spatiaux ou temporels surlignés, ou au contraires des termes non surlignés), néanmoins cela fournit une aide notable à l'assesseur. L'information spatiale étant plus complexe à évaluer sans suffisamment de connaissances, une carte a été ajoutée pour localiser toutes les informations spatiales du document ou faire des recherches. Cette carte est située au centre de l'interface (figure 7.4). Pour effectuer le jugement, l'assesseur remplit un tableau dans lequel il y a une case à cocher pour chaque facette ainsi que

pour l'ensemble de la requête. De plus l'interface lui indique le niveau d'avancement de chaque requête. Par exemple, sur la figure 7.4, nous pouvons voir que l'assesseur a jugé 278 documents sur les 300 associés à cette requête.

Cette plateforme a été conçue en (X)HTML, CSS et PHP associés à une base de données MySQL. La base de données permet de stocker les différentes requêtes à évaluer et la liste des documents résultats associés à chacune, ainsi que les comptes des assesseurs et les affectations des requêtes. La carte utilisée est de type GoogleMaps et les informations spatiales qui y sont affichées ont été converties en KML (voir section 2.3.1 page 25) grâce aux fonctions disponibles dans PostGIS.

PIVasse a permis de constituer notre base de référence pour évaluer nos approches de combinaison et les comparer. Cette plateforme peut être utilisée pour évaluer d'autres corpus : elle n'est ni spécifique à notre corpus ni à notre SRI.

7.6 Conclusion

Dans ce chapitre nous avons présenté les différents prototypes développés pour tester et vérifier nos hypothèses. L'ensemble des ces prototypes est disponible sur notre site [34].

PIV^2 est un prototype uniformisant des index spatiaux et temporels existants (en l'occurrence ceux générés par la plateforme PIV). Il crée ainsi de nouveaux index composés de tuilages de grain différents. PIV^2 pourrait uniformiser des index issus d'autres SRIG.

PIVcomb a été mis en place pour mettre en œuvre la combinaison par contraintes. Il permet d'interroger différents SRI, voire plusieurs fois le même, et d'associer des contraintes (exigences, préférences, opérateurs) à chaque critère.

Concernant l'évaluation d'un SRIG, nous avons mis en place deux plateformes Web. La première, PIVone, nous a permis d'interroger les différents index (de PIV ou PIV^2), de vérifier la répartition des résultats pour chaque facette et pour la combinaison. Il permet également de récupérer le *pool* de documents correspondants à l'ensemble des résultats. La deuxième plateforme, PIVasse, a été mise en place pour constituer une base de référence permettant à des assesseurs de juger les différents résultats de chaque requête. L'objectif est de faciliter au maximum la tâche des assesseurs qui est fastidieuse. Pour cela l'interface propose le surlignement des informations spatiales et temporelles ou encore la localisation des informations spatiales sur une carte.

Dans le chapitre qui suit, nous allons présenter les différentes expérimentations réalisées ainsi que les résultats qui en ont découlé.

34. http://t2i.univ-pau.fr/protos/

Chapitre 8

Expérimentations

Sommaire

8.1 Introduction

Pour vérifier nos différentes propositions, nous avons mis en place plusieurs expérimentations. Tout d'abord, nous avons évalué l'intérêt de notre approche d'uniformisation notamment concernant la perte de précision induite par le tuilage. Pour cela nous avons comparé un SRI spatial classique (PIV) avec un SRI spatial avec uniformisation (PIV^2). Une première expérimentation a ainsi été réalisée pour l'information spatiale, puis une autre pour l'information temporelle. Ensuite nous avons évalué l'intérêt de la combinaison de SRI spécialisés (spatial, temporel et thématique). Nous avons ainsi testé plusieurs combinateurs linéaires dans un premier temps, puis notre approche de combinaison par contraintes (PIVcomb).

Notre corpus actuel contient 11 livres, ce qui correspond à 5645 paragraphes, 8983 ES (7608 ESA[35] et 1375 ESR[36]) et 1702 ET (1561 ETA[37] et 141 ETR[38]). Chaque paragraphe peut contenir entre zéro et plusieurs ES (idem pour les ET).

8.2 Évaluation de l'approche d'uniformisation appliquée à l'information spatiale

Concernant l'uniformisation spatiale, des expérimentations ont été mises en place pour répondre à plusieurs interrogations :

- globalement l'approche d'uniformisation a-t-elle un intérêt ? N'engendre-t-elle pas une perte de précision trop importante pouvant dégrader très fortement la qualité des résultats retournés par un SRI spatial ?
- par rapport à notre corpus, un tuilage et une formule de pondération se démarquent-ils ?
- selon le type de relation spatiale, faut il utiliser une formule de pondération différente ?
- utiliser un index adapté à la granularité de la requête améliore-t-il les résultats ?

Ces expérimentations ont donné lieu à deux publications : une dans *Workshop on Geographic Information on the Internet* (WGII) associé à la conférence ECIR'09 [PSG09] et une dans *14th International Symposium on Spatial Data Handling* [PSG10].

L'expérimentation présentée ici porte sur un échantillon de notre corpus : 1019 paragraphes ce qui correspond à 1028 ES (902 ESA et 126 ESR) et 40 requêtes. 15 requêtes portent sur une ESA : 5 de chaque type (granularité petite telle que pic, granularité intermédiaire telle que commune, granularité large telle que région). 25 portent sur une ESR : 5 de chaque type (orientation, proximité, union, inclusion, distance). Nous avons choisi de travailler sur un tel échantillon afin de pouvoir évaluer tous les documents

35. ESA : Entité Spatiale Absolue
36. ESR : Entité Spatiale Relative
37. ETA : Entité Temporelle Absolue
38. ETR : Entité Temporelle Relative.

pour chaque requête dans l'optique de pouvoir calculer la précision et le rappel. Trois personnes ont jugé l'ensemble des documents pour chaque requête. Nous allons donc présenter maintenant les analyses permettant de répondre aux trois interrogations que nous avons listées.

8.2.1 Comparaison des SRI spatiaux PIV et PIV^2

La première analyse consiste à évaluer l'intérêt de l'uniformisation spatiale, notamment vérifier que la perte pouvant être induite par l'utilisation d'un tuilage n'est pas trop importante. Pour cela, nous avons comparé notre système de RI spatial uniformisé, PIV^2, avec le système de RI spatial PIV existant. Comme nous l'avons vu dans la section 7.3, PIV génère un index stockant toutes les entités spatiales identifiées et validées. Ces ES peuvent être de différents types (absolues/relatives) et de différentes granularités (communes, départements, pics, ...). Le calcul de similarité, entre une information spatiale d'un document (Df) et d'une requête (Q) (figure 8.1), est basé sur la surface de recouvrement entre les représentations de ces deux informations (I) ainsi que sur la distance entre le centre de cette zone I et le centre de la boîte englobante représentant la requête (voir formule 8.1) [SGPL08].

$$score(Df) = \frac{\frac{I}{Df} + \frac{I}{Q}}{2 + \frac{d}{D}} \tag{8.1}$$

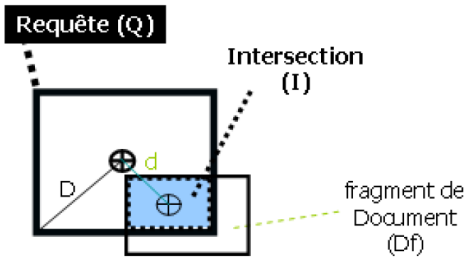


FIGURE 8.1 – Calcul de pertinence d'une ES d'un document pour une requête donnée dans le système PIV (illustration extraite de [SGPL08])

Nous avons donc comparé PIV à PIV^2 intégrant notre approche d'uniformisation spatiale présentée dans le chapitre 5. Étant donné que nous allons traiter la comparaison entre les différents tuilages et formules de pondération dans la section suivante, nous ne comparons ici que le meilleur couple (tuilage, pondération) de PIV^2 avec l'approche utilisée par PIV. Comme nous pouvons le voir dans le tableau 8.1, PIV^2 donne des résultats supérieurs de 13 % à PIV. Le test `t` de Student pairé (voir section 6.4.2) entre les résultats de PIV et PIV^2 est égal à 0,02, donc inférieur à 0,05, ce qui implique que la différence entre les deux est significative. Donc, la perte de précision engendrée par

le tuilage est compensée par l'utilisation de calculs de fréquences pour la pondération puisqu'elle permet d'améliorer significativement les résultats.

SRIG	MAP	Amélioration
PIV (baseline)	0,62	0 %
PIV^2	0,70	+13 %

TABLE 8.1 – Comparaison PIV - PIV^2 (meilleur tuilage spatial et formule de pondération)

8.2.2 Analyse et comparaison de différents tuilages spatiaux et formules de pondération

La deuxième analyse porte sur la comparaison de différents tuilages possibles ainsi que sur les différentes formules de pondération utilisables. Nous avons testé six tuilages : trois administratifs (communal, départemental et régional) et trois réguliers (100×100,200×200 et 400×400). La superficie des tuiles de la grille 200×200 correspond à peu près à celles du tuilage communal. Pour la pondération, nous avons utilisé les quatre formules présentées dans la section 5.3 (TF, TF·IDF, OkapiBM25, TFp). Comme nous pouvons le voir dans le tableau 8.2, le tuilage communal associé à la formule de pondération TFp donne les meilleurs résultats. La formule TFp s'avère de plus donner les meilleurs résultats pour chacun des tuilages testés.

Tuilage	TF	TF·IDF	OkapiBM25	TFp
Communal	0,61	0,61	0,63	**0,70**
Départemental	0,40	0,39	0,40	**0,53**
Régional	0,40	0,39	0,39	**0,56**
Grille de 100x100	0,59	0,59	0,62	**0,68**
Grille de 200x200	0,61	0,60	0,63	**0,69**
Grille de 400x400	0,63	0,62	0,65	**0,66**

TABLE 8.2 – Comparaison de différents tuilages spatiaux et formules de pondération (MAP)

8.2.3 Analyse par type de relation spatiale

Ensuite nous avons voulu tester si selon le type de relation spatiale, une formule de pondération est plus adaptée. Suite aux résultats de l'expérimentation précédente, la comparaison s'effectue sur un tuilage communal. Le tableau 8.3 montre que pour chaque type de relation spatiale que nous traitons, le TFp donne les meilleurs résultats.

8.2.4 Test de l'index de granularité la plus proche de celle de la requête

Enfin, nous avons voulu vérifier l'intérêt d'utiliser un tuilage multi-échelles présenté dans la section 5.2.2. Il s'agit de générer plusieurs index à des échelles différentes, pour

Relation Spatiale	TF	TF·IDF	OkapiBM25	TFp
Distance	0,52	0,52	0,57	**0,69**
Inclusion	0,69	0,69	0,69	**0,76**
Orientation	0,69	0,68	**0,70**	**0,70**
Proximité	0,53	0,53	0,54	**0,69**
Union	0,64	0,63	0,66	**0,78**

TABLE 8.3 – Comparaison des différentes formules de pondération sur un tuilage communal pour chaque type de relation spatiale (MAP)

le spatial par exemple ici : communal, départemental et régional ; puis, lors de l'interrogation, de déterminer la granularité de la requête, et, d'utiliser l'index le plus adapté à cette granularité. Ainsi si l'utilisateur soumet une requête dont la granularité est une commune (telle que « à Pau »), l'index de tuilage communal sera utilisé.

Pour évaluer cette approche nous avons sélectionné 10 requêtes (cinq portant sur une commune vu que le tuilage donnant les meilleurs résultats est le tuilage communal et cinq portant sur un tuilage différent) parmi les 15 portants sur une ESA. Nous avons donc comparé cette approche multi-échelles à l'utilisation d'un seul index, celui choisi par défaut par rapport au contenu du corpus (pour le spatial il s'agit du tuilage communal).

Comme nous pouvons le voir dans le tableau 8.4, utiliser un tuilage multi-échelles améliore les résultats du SRI : 8,5 % d'amélioration par rapport au meilleur tuilage (communal), ce qui n'est néanmoins pas significatif : test `t` de Student pairé $p = 0{,}11 > 0{,}05$ (voir section 6.4.2). Il faut noter que l'utilisation d'un index adapté à la granularité de la requête va par contre améliorer les performances, en termes de rapidité, d'interrogation du SRI. En effet, pour le tuilage communal, la table contenant les liens tuiles-documents a plus de 5,8 million d'entrées contre 8700 pour le tuilage régional. Ceci est une problématique connue et étudiée dans les SIG d'où des approches telles que les *quadtree* [Sam84].

Approche	MAP	Amélioration
Tuilage par défaut (communal)	0,60	0,0 %
Tuilage adapté à la requête (multi-échelles)	0,65	+8,5 %

TABLE 8.4 – Comparaison de l'approche multi-échelles au tuilage par défaut

En conclusion, l'uniformisation spatiale couplée à l'usage de formules classiques de RI améliore les résultats de l'approche de RI spatiale classique PIV. L'utilisation d'un tuilage implique une perte de précision. Néanmoins, l'utilisation de fréquences d'apparition des tuiles permet de compenser cette perte. De plus, calculer des poids en tenant compte de l'ensemble des tuiles d'une unité documentaire permet une première prise en compte du contexte. Pour notre corpus, le couple tuilage communal – TFp donne les meilleurs résultats. Si nous regardons la répartition du type d'information spatiale que contient notre corpus (voir figure 8.2), nous remarquons que la majorité des informations se distingue du cadre administratif (pics, lacs, . . .) et donc ne correspondant à aucun tuilage administratif. Si par contre nous regardons uniquement la répartition des ES de type

administratif (voir figure 8.3), nous remarquons que 70% sont des communes, ce qui peut expliquer que le tuilage communal donne les meilleurs résultats.

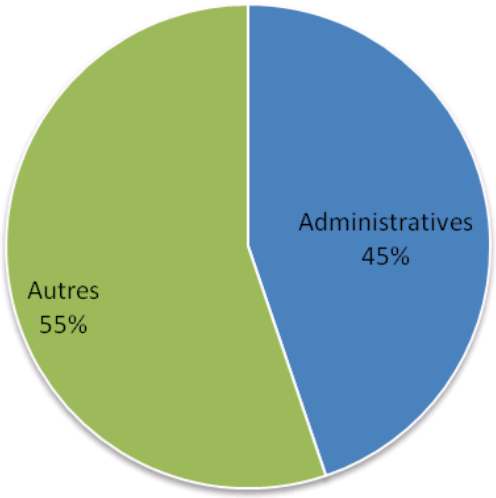


FIGURE 8.2 – Répartition des ES dans notre corpus

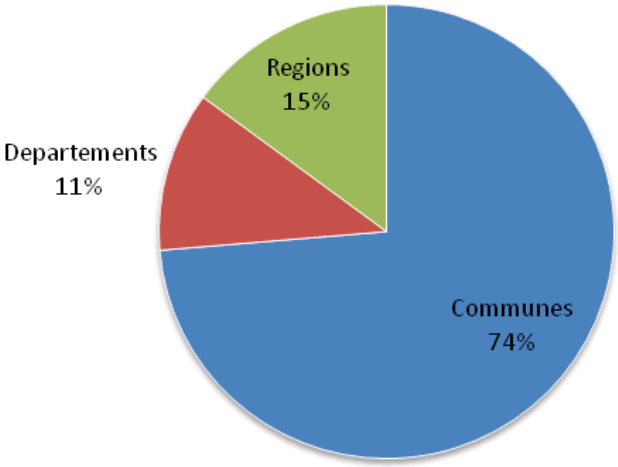


FIGURE 8.3 – Répartition des ES administratives dans notre corpus

8.3 Évaluation de l'approche d'uniformisation appliquée à l'information temporelle

Pour l'uniformisation temporelle, des expérimentations ont été mises en place pour répondre à plusieurs interrogations :

- l'approche d'uniformisation appliquée à l'information temporelle donne-t-elle de bons résultats comme pour l'information spatiale ?
- par rapport à notre corpus, un tuilage et une formule de pondération se démarquent ils ?

Contrairement au spatial, nous avons utilisé la totalité de notre corpus : les 11 livres soit 5645 paragraphes (1702 ET : 1561 ETA et 141 ETR). Nous avons soumis 35 requêtes à PIV et PIV^2. 15 requêtes portent sur une ETA : 5 de chaque type (granularité petite telle que jour, granularité intermédiaire telle que mois, granularité large telle que année). 20 portent sur une ETR : 5 de chaque type (orientation, proximité, intervalle, inclusion). Vu la quantité d'unités documentaires à juger (5645x35 soit environ 200000), nous avons utilisé une approche de type *pooling* telle que décrite dans la section 6.4.1. Pour obtenir le *pool* de documents nous avons utilisé les 100 premiers documents restitués par chacun des systèmes que nous avons comparé (PIV, PIV^2 avec six tuilages différents soit sept SRI). Nous allons donc présenter maintenant les analyses permettant de répondre aux deux interrogations que nous avons listées.

8.3.1 Comparaison des SRI temporels PIV et PIV^2

La première analyse consiste à évaluer l'uniformisation temporelle. Comme pour l'information spatiale, nous avons comparé notre système de RI temporel uniformisé, PIV^2, avec le système de RI temporel PIV existant. Comme nous l'avons vu dans la section 7.3, PIV génère un index stockant toutes les entités temporelles identifiées et validées. Ces ET peuvent être de différents types (absolues/relatives) et de différentes granularités (jours, mois, années, ...). Le calcul de similarité, entre une information temporelle d'un document (Df) et d'une requête (Q) (figure 8.4), est basé sur la surface recouvrement entre les représentations (durées) de ces deux informations (I) ainsi que sur la distance entre le centre de cet intervalle I et le centre de l'intervalle représentant la requête (voir formule 8.2) [LGS07].

$$score(Df) = \frac{\frac{I}{Df} + \frac{I}{Q}}{2 + \frac{d}{D}} \tag{8.2}$$

Nous avons donc comparé PIV à PIV^2 intégrant notre approche d'uniformisation temporelle présentée dans le chapitre 5. Étant donné que nous allons traiter la comparaison entre les différents tuilages et formules de pondération dans la section suivante, nous ne comparons ici que le meilleur couple (tuilage, pondération) de PIV^2 avec l'approche utilisée par PIV. Comme nous pouvons le voir dans le tableau 8.5, contrairement au spatial, l'uniformisation temporelle donne des résultats légèrement en-dessous à ceux de PIV (0,5 % de moins). Le test `t` de Student pairé (voir section 6.4.2) entre PIV et

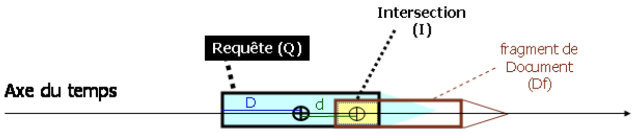


FIGURE 8.4 – Calcul de pertinence d'une ET d'un document pour une requête donnée dans le système PIV (illustration extraite de [LGS07])

PIV^2 est compris entre 0,47 et 0,80 selon la formule de pondération utilisée, donc non inférieur à 0,05, ainsi la différence entre PIV^2 et PIV n'est pas significative. Contrairement au spatial, les informations temporelles sont peu nombreuses, et donc rares sont les paragraphes à contenir plusieurs informations temporelles. Ceci explique en effet que l'utilisation de calculs de fréquences pour la pondération n'améliore pas les résultats.

SRIG	MAP	Amélioration
PIV (baseline)	0,93	0,0 %
PIV^2	0,92	−0,5 %

TABLE 8.5 – Comparaison PIV - PIV^2 (meilleur tuilage temporel et formule de pondération)

8.3.2 Analyse et comparaison de tuilages temporels et formules de pondération

La deuxième analyse porte sur la comparaison de différents tuilages possibles ainsi que sur les différentes formules de pondérations utilisables. Nous avons testé six tuilages : trois calendaires (semaines, mois et années) et trois réguliers (grilles de 10, 30 et 400 jours). Ici aussi nous avons utilisé les quatre formules de pondération présentées dans la section 5.3 (TF, TF·IDF, OkapiBM25, TFp). Comme nous pouvons le voir dans le tableau 8.6, le tuilage mensuel est celui qui donne les meilleurs résultats. Par contre, concernant la formule de pondération, aucune ne se dégage réellement. Les informations temporelles étant très peu nombreuses dans chaque paragraphe (0 ou 1 ET en général), le TFp qui pondère différemment chaque ET d'un même paragraphe ne produit donc pas de résultats très différents.

En conclusion, l'uniformisation temporelle donne des bons résultats très légèrement inférieurs à ceux de notre référence (PIV). Même si ici l'utilisation de la fréquence des tuiles a un impact faible du fait que rares sont les paragraphes à contenir plusieurs ET, la perte de précision due à l'uniformisation reste limitée. Concernant la formule de pondération aucune ne se démarque. Le tuilage donnant les meilleurs résultats est le tuilage mensuel. Contrairement au spatial, la répartition des informations temporelles est plus diversifiée dans notre corpus (voir figure 8.5) : essentiellement des années (47 %), des jours (35 %) et des mois (10 %). A priori le meilleur tuilage devrait être le tuilage annuel

Tuilages	TF	TF-IDF	OkapiBM25	TFp
Hebdomadaire	0,89	**0,90**	0,89	0,85
Mensuel	**0,92**	**0,92**	**0,92**	**0,92**
Annuel	0,83	0,83	0,83	**0,87**
Grille de 10 jours	0,89	**0,90**	0,89	0,89
Grille de 30 jours	0,87	0,87	**0,88**	**0,88**
Grille de 400 jours	0,74	0,74	**0,76**	0,72

TABLE 8.6 – Comparaison de différents tuilages temporels et formules de pondération (MAP)

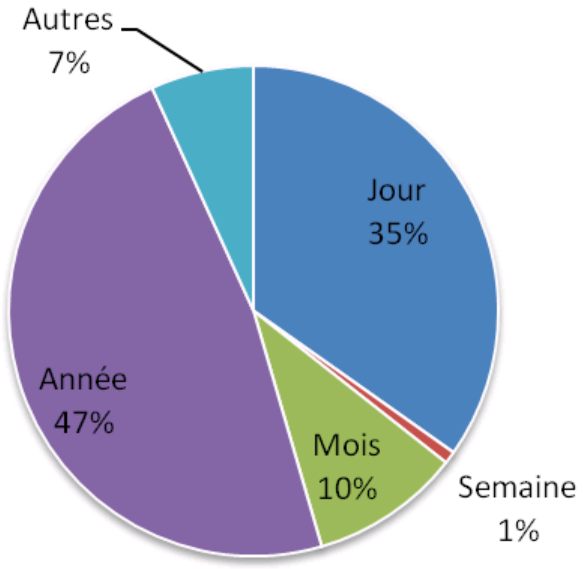


FIGURE 8.5 – Répartition des ET calendaires dans notre corpus

mais étant donné qu'il a y aussi beaucoup d'informations temporelles correspondantes à des jours, le tuilage mensuel semble être le meilleur compromis.

8.4 Évaluation de l'approche par combinaison appliquée à l'information géographique

Pour réaliser nos expérimentations sur la combinaison, nous nous sommes basés sur le cadre d'évaluation d'un SRI géographique que nous avons proposé, étant donné qu'il n'en existe pas actuellement. Nous avons donc appliqué ce cadre au prototype PIV^2 et évalué l'apport de la combinaison de critères [PCSH10a, PCSH10b].

Dans un premier temps, nous avons donc mis en place une collection de test, nommée MIDR_2010. Puis nous avons comparé les différents opérateurs linéaires existants pré-

sentés dans la section 6.2 afin de vérifier que CombMNZ donne les meilleurs résultats en RIG tout comme l'a montré Lee pour la RI [Lee97]. Ensuite nous avons étudié l'apport de chaque facette géographique ainsi que leur complémentarité. Pour finir, nous avons comparé notre approche de combinaison par contraintes (PIVcomb) avec CombMNZ.

8.4.1 Mise en place de la collection de test MIDR_2010

La collection de test MIDR_2010[39] comprend les quatre volets identifiés en section 6.4.1.

Premièrement, comme précédemment, 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 restitué à l'usager par le SRI géographique est un de ces paragraphes, vu comme le meilleur point d'entrée dans l'ouvrage associé.

Deuxièmement, 34 *topics* couvrant tout ou partie des trois facettes de l'information géographique ont été constitués.

Troisièmement, les *qrels* ont été obtenus en interrogeant trois SRI – un thématique basé sur le modèle PL2 (configuration de base de Terrier), un spatial et un temporel – avec le titre des topics. Pour le spatial et le temporel, il s'agit de PIV^2 mettant en place l'uniformisation. Le tuilage communal et le tuilage mensuel ont été retenus suite aux résultats des expérimentations présentées juste auparavant.

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 facette ainsi qu'un jugement booléen global, ces quatre évaluations étant agrégées pour constituer un jugement graduel, comme expliqué en section 6.4.1. Ces jugements ont été réalisés par 12 assesseurs bénévoles.

8.4.2 Comparaison des opérateurs linéaires

Lee a montré que CombMNZ donne les meilleurs résultats en RI classique [Lee97]. Nous avons voulu vérifier qu'il en est de même pour la collection MIDR_2010, en confrontant les performances respectives des cinq combinateurs présentés dans 6.2 : CombMAX, CombMIN, CombSUM, CombMNZ et CombANZ.

Le tableau 8.7 montre les performances des combinateurs appliqués aux trois facettes (thématique, spatiale et temporelle) selon la mesure $NDCG$. Ces derniers sont déclinés en version normalisée et non normalisée. L'effet de la normalisation, inférieur à 6 %, est également indiqué. Cette expérimentation menée sur la collection MIDR_2010 corrobore les observations de Lee [Lee97] : elle montre que la configuration optimale (0,7887) correspond à CombMNZ normalisé.

39. La collection de test est accessible sur http://t2i.univ-pau.fr/MIDR/.

	Combinateurs Comb*				
	Min	Max	Sum	Anz	Mnz
Non normalisée	0,5909	0,7418	0,7706	0,7289	0,7788
Normalisée	0,6222	0,7458	0,7729	0,7237	**0,7887**
Gain (%)	5,30	0,54	0,30	−0,71	1,27

TABLE 8.7 – Performances relatives de combinateurs et effet de la normalisation.

Dans la section suivante, nous testons la performance de la combinaison basée sur CombMNZ normalisé, en fonction des facettes utilisées.

8.4.3 Analyse comparative de la performance des différentes combinaisons de critères spatiaux, temporels et thématiques mises en œuvre avec CombMNZ

Le tableau 8.8 présente les comparaisons effectuées entre les différents SRI et deux « baselines » thématiques (SRI de référence) identifiées dans [POGCGVUL08] : 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 6.4.2). Calculés avec le programme trec_eval [40] version 9.0 utilisé à Trec, ces résultats portent sur l'analyse des performances des moteurs au regard des 34 requêtes.

Globalement, il est intéressant de noter que ces deux baselines sont quasiment identiques. Contrairement aux résultats rapportés par Perea-Ortega et al [POGCGVUL08], 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 documents de tailles assez homogènes (*versus* des documents entiers de tailles très 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 monocritères, on observe une performance maximale de 0,4846 pour le SRI spatial. Ces SRI sont toutefois caractérisés par des performances similaires. Par ailleurs, le gain apporté par la combinaison de facettes hétérogènes est observable à partir de deux facettes combinées. Notons que ce gain est statistiquement significatif au regard des deux baselines. Nous remarquons que l'alliance du spatial et du temporel fournit les meilleures performances (0,7114). Toutefois, la combinaison du thématique et du temporel offre une performance similaire (0,6991). Ce résultat 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 facettes apporte la performance maximale (0,7887). Le gain de précision correspondant à 66,4 % par rapport à la baseline forte Th^+ valide l'hypothèse à la base de notre travail : **combiner les trois facettes de l'information**

40. http://trec.nist.gov/trec_eval/

Fusion de NSRI	SRI monocritères				$moyNDCG$	Amélioration (%)	
	Th$^-$	Th$^+$	S	Te		Th$^-$	Th$^+$
1	✓				0,4742	0,0	0,0
		✓			0,4741	0,0	0,0
			✓		0,4846	2,2	2,2
				✓	0,4810	1,4	1,5
2	✓	✓			0,4742	0,0	0,0
	✓		✓		0,6307*†	33,0	33,0
	✓			✓	0,6986*†	47,3	47,4
		✓	✓		0,6310*†	33,1	33,1
		✓		✓	0,6991*†	47,4	47,5
			✓	✓	0,7114*†	50,0	50,1
3	✓	✓	✓		0,6250*†	31,8	31,8
	✓	✓		✓	0,6819*†	43,8	43,8
	✓		✓	✓	0,7872*†	66,0	66,0
		✓	✓	✓	**0,7887***†	**66,3**	**66,4**
4	✓	✓	✓	✓	0,7580*†	59,8	59,9

TABLE 8.8 – Efficacité des SRI par rapport aux baselines thématiques. Le symbole '*' (resp. †) indique une différence significative par rapport à la baseline Th$^-$ (resp. Th$^+$).

géographique améliore la pertinence des résultats par rapport à la facette thématique seule.

En complément, une analyse approfondie montre que combiner les trois facettes offre des résultats significativement supérieurs à la meilleure combinaison de deux facettes (S+Te) : différence de 10,9 % entre les moyNDCG (0,7887 pour les 3 facettes et 0,7114 pour S+Te, voir tableau 8.8), test t de Student pairé $p < 0{,}001$ (voir section 6.4.2). Notons qu'une combinaison des trois facettes avec les deux versions thématiques (Th$^+$ et Th$^-$) n'apporte pas d'amélioration supplémentaire (0,7580) (voir tableau 8.8), le renforcement thématique masquant alors les aspects complémentaires apportés par les deux autres facettes.

8.4.4 Analyse par topic de la combinaison linéaire CombMNZ

Après avoir identifié la combinaison de facettes optimisant la performance de PIV^2 (Th$^+$+S+Te), nous analysons dans cette section sa performance par topic. Cette analyse plus détaillée a pour objectif d'affiner les observations précédentes. À cet effet, le tableau 8.9 synthétise, pour chaque topic :

1. le numéro du topic (colonne 1),
2. le nombre de documents jugés pertinents (cf. *qrels*) pour chacune des trois facettes (colonnes 2 à 4),

3. la performance de PIV[2] pour le topic considéré (colonne 5),
4. et le nombre de documents pertinents restitués par un seul, par uniquement deux ou par les trois SRI, chacun étant associé à une facette (colonnes 6 à 8).

La répartition des documents pertinents par facette est variable d'un topic à l'autre. On remarque une prédominance globale de la facette spatiale. Ceci est dû à la spécificité des documents de la collection MIDR_2010 qui traitent de sujets liés à la zone spatiale correspondant aux Pyrénées. Par ailleurs, l'écart-type faible $\sigma_{NDCG} = 0{,}0621$ indique une performance par topic stable. Globalement, la différence de performance par rapport à la moyenne de 0,7887 calculée sur les 34 topics est dans $[-24\,\%\,;\,+13\,\%]$. Enfin, il est intéressant d'étudier la contribution des trois facettes : sont-elles redondantes (restituant des documents similaires) ou bien complémentaires (restituant des documents spécifiques à chaque facette) ? La dernière colonne du tableau 8.9 montre un faible nombre de documents simultanément restitués par les trois facettes, démontrant par conséquent que **les trois facettes sont complémentaires**.

8.4.5 Comparaison de la combinaison linéaire CombMNZ avec la combinaison par contraintes PIVComb

La dernière expérimentation réalisée avait pour objectif de vérifier l'apport de notre approche par contrainte. Pour cela, nous avons comparé le système PIVcomb mettant en œuvre cette approche avec la meilleure des combinaisons linéaires standards, CombMNZ (voir section précédente). La collection et la démarche utilisées sont donc similaires à celles présentées dans la section 8.4.1 avec 10 *topics* (les seuls à être contraints). Nous avons aussi comparé notre combinaison par contraintes étendue aux deux approches de combinaisons par priorités (*Prioritized Scoring Model*, et, *Prioritized And Model*) de Costa Pereira et al [CPDP09a] afin de déterminer si elles peuvent être utilisées pour la RIG (voir section 6.3).

Comme nous pouvons le voir dans le tableau 8.10, les approches de combinaison par priorités donnent des résultats moins bons que CombMNZ. Pour l'approche *Prioritized And Model*, cette perte est due au fait que tous les critères sont obligatoires. Or, dans certaines requêtes, des critères sont optionnels et donc cette approche ne permet pas de retourner les résultats ne satisfaisant pas ces critères optionnels. Concernant l'approche *Prioritized Scoring Model*, seul un critère est obligatoire, donc elle retourne des résultats non pertinents pour des requêtes où deux critères ou plus sont exigés. Il faut néanmoins noter que ces approches n'ont pas été proposés pour ce contexte particulier et donc ne sont pas adaptées à des requêtes aussi complexes.

L'approche de combinaison étendue via les contraintes permet de réaliser une recherche plus précise en spécifiant une exigence, une préférence et un opérateur par critère. Elle permet de plus de gérer des scénarios de recherches, non possibles avec les autres opérateurs, tel que l'exclusion d'un critère. Ce sont les raisons pour lesquelles nous observons un gain de précision de 54,16 % par rapport à CombMNZ. Il faut néanmoins noter que ces différentes approches n'ont pas servi à la constitution du *pool*, or,

Topic	Nombre de documents pertinents			$NDCG$	Complémentarité		
	Th^+	S	Te		1 crit	2 crit	3 crit
1	67	260	134	0,7715	148	134	15
2	60	213	138	0,8756	195	87	14
3	42	111	101	0,7808	201	25	1
4	16	16	74	0,5960	82	9	2
5	94	87	34	0,8169	144	34	2
6	48	60	59	0,8496	139	14	2
7	71	24	93	0,7354	127	29	1
8	27	177	103	0,8098	229	39	1
9	41	209	112	0,7949	154	95	6
10	68	110	102	0,7703	176	52	6
11	33	111	6	0,7137	90	30	6
12	41	246	116	0,7638	150	101	17
13	106	120	47	0,7683	201	33	2
14	74	110	107	0,7857	137	74	2
15	89	152	103	0,7010	190	74	2
16	74	40		0,8072	84	15	2
17	197	116	23	0,8315	190	70	2
18	37	141	112	0,8758	218	36	2
19	113	175	101	0,7163	199	83	8
20	180	11	73	0,7327	147	57	1
21	13	150	122	0,7083	157	58	4
22	7	162	121	0,8882	198	46	4
23	126	114	120	0,8360	205	58	13
24	170	71	30	0,8160	185	41	1
25	234	102	110	0,8195	132	100	38
26	160	227	131	0,7918	71	126	65
27	19	41	113	0,8580	149	12	65
28	21	94	117	0,8326	139	45	1
29	47	41	120	0,7749	149	25	3
30	51	91	131	0,8751	213	30	3
31	16	194	55	0,8496	185	40	3
32	5	224	36	0,7510	160	93	2
33	23	60	37	0,7552	84	4	2
34	170	253	120	0,7627	95	116	72
Moyenne	74,7	126,9	90,9	0,7887	156,6	55,4	10,9

TABLE 8.9 – Étude par topic de la distribution des documents pertinents selon les trois facettes, de la performance du SRI PIV^2 et de la complémentarité des facettes.

Approche	MAP	Gain(%)
CombMNZ	0,1658	0,0
Prioritized And Model	0,0705	−57,48
Prioritized Scoring Model	0,1034	−37,64
PIVcomb	0,2556	54,16

TABLE 8.10 – Comparaison de différentes approches de combinaison

les résultats non jugés sont considérés comme non pertinents, ce qui est une des raisons expliquant les faibles valeurs des MAP.

Cette première expérimentation reste limitée, il n'y a que 10 *topics*, or il est nécessaire d'en évaluer au minimum 25 (voir section 6.4.2). Toutefois, les résultats encourageants (+54,16% par rapport à CombMNZ) montrent l'intérêt de cette approche. Nous avons prévu à court terme de réaliser une expérimentation plus conséquente pour valider ces résultats.

8.5 Conclusion

L'approche d'uniformisation présentée dans le chapitre 5 a pour but d'homogénéiser les méthodes d'indexation et de RI de chacun des critères à combiner. Il était donc nécessaire de vérifier que notre approche d'uniformisation ne dégrade pas les résultats de chacun des SRI de manière conséquente. Les expérimentations réalisées ont confirmé l'intérêt de notre approche d'uniformisation appliquée à l'information spatiale (résultats améliorés par rapport à notre référence PIV) et à l'information temporelle (résultats similaires à notre référence PIV).

Nous avons émis l'hypothèse que la combinaison des différentes facettes de l'information géographique améliore la pertinence des résultats. Les expérimentations réalisées, comparant notamment un SRI classique (Terrier) avec PIV^2 combinant l'uniformisation spatiale, temporel et Terrier, confirment cette hypothèse. Elles permettent de plus de quantifier cet apport : nous observons une amélioration de 66,4 %, qui plus est significative. De plus, les trois facettes ne sont pas redondantes, mais bien complémentaires.

Ensuite nous avons proposé une approche de combinaison étendue via des contraintes que nous avons comparée à la meilleure combinaison linéaire standard (CombMNZ) et aux approches de combinaison par priorité. Notre approche se révèle la plus performante (+54,16 %).

Les expérimentations réalisées ici présentent plusieurs limites. Premièrement, de par ses 5 645 paragraphes totalisant 3,7 Mo, le corpus utilisé est très peu volumineux en comparaison des collections TREC. Deuxièmement, le nombre de requêtes utilisées a été en règle générale compris entre 30 et 40. Cette valeur représente davantage que le minimum de topics à considérer pour pouvoir réaliser des analyses statistiques valides (qui est de 25) mais reste en deçà des 50 requises dans TREC. Nous continuons l'effort de jugement manuel des documents permettant d'obtenir davantage de topics à analyser.

Conclusion de la contribution

Nous avons présenté les différentes contributions de ce travail y compris les développements et expérimentations effectués.

Le chapitre 5 a présenté notre approche d'uniformisation générique permettant d'homogénéiser différents critères portant sur des informations de une à n dimensions. Nous nous sommes néanmoins limités aux deux premières dimensions : nous avons appliqué cette approche à l'information temporelle (calendaire – une dimension) et spatiale (cartographique – deux dimensions). Pour l'indexation, cette approche consiste à uniformiser les différents types de représentations d'informations en un type unique : la tuile. Plusieurs types de tuilages sont possibles : régulier ou explicite (calendaire pour le temporel, administratif pour le spatial). À chaque tuile est associé un poids lié à sa fréquence d'apparition, cette dernière pouvant être binaire (0 ou 1) ou proportionnelle (comprise entre 0 et 1). Nous avons appliqué les formules utilisées couramment en RI (TF, TF·IDF, OkapiBM25) sur des fréquences binaires et nous avons utilisé le TF sur des fréquences proportionnelles (TFp). Pour l'interrogation, nous avons utilisé le modèle vectoriel de Salton et le produit scalaire. L'information pouvant être exprimée sous différentes échelles dans les textes, nous avons de plus proposé d'utiliser un tuilage multi-échelle pour générer des index de granularités différentes puis utiliser l'index le plus adapté au grain de la requête de l'utilisateur.

Le chapitre 6 a présenté tout d'abord plusieurs approches de combinaison linéaire standards permettant de tester rapidement la combinaison de différentes facettes de l'information géographique (spatial, temporel et thématique) afin d'évaluer l'intérêt de cette combinaison et de quantifier son apport. Néanmoins, avec ces approches, l'utilisateur ne peut pas personnaliser la combinaison. Nous proposons alors une approche de combinaison étendue par contraintes permettant à un utilisateur de paramétrer la manière dont sera traité chaque critère (obligatoire/optionnel/ à exclure ?, quelle est son importance ?, quel opérateur ?) et ainsi la combinaison. Enfin, nous proposons un cadre d'évaluation d'un SRI géographique nous permettant notamment de comparer ces différentes approches de combinaison.

Le chapitre 7 a exposé les différents prototypes mis au point durant ce travail de thèse. PIV^2 permet d'uniformiser des données spatiales et temporelles déjà indexées. Il génère ainsi de nouveaux index constitués de tuiles. PIVcomb sert à combiner différents critères géographiques d'interrogation en spécifiant pour chacun les exigences, préférences et opérateurs. Concernant l'évaluation des SRIG, nous avons mis en place

deux plateformes Web. La première, PIVone, permet d'interroger les différents SRI, de vérifier qu'il y a suffisamment de résultats sur chacun des critères et de récupérer le *pool* de document associé à chacune des requêtes. La deuxième, PIVasse, a été mise en place pour faciliter le travail des assesseurs (surlignement des informations spatiales et temporelles, carte, ...) consistant à juger les différents documents potentiellement pertinents pour chaque requête.

Le dernier chapitre (chapitre 8) a détaillé nos différentes expérimentations. Elles démontrent l'intérêt de notre approche d'uniformisation : les résultats sont meilleurs que notre référence (PIV) pour le spatial et similaires à notre référence pour le temporel. Concernant la combinaison de différents critères géographiques, nous avons testé plusieurs approches de combinaison linéaire standard : CombMNZ est celle qui donne les meilleurs résultats. L'expérimentation réalisée a, de plus, montré le fort apport de la combinaison des facettes spatiales et temporelles par rapport aux termes seuls : 66,4 % d'amélioration, qui est significative. Cette étude montre aussi que ces facettes ne sont pas redondantes mais bien complémentaires. Concernant la combinaison par contraintes, elle permet d'exprimer beaucoup plus précisément les requêtes et, par conséquent, donne de meilleurs résultats que CombMNZ (54,16 % d'amélioration).

Nous avons choisi de traiter chaque facette de l'information géographique de manière spécifique et indépendante comme le préconisent de nombreux travaux en GIR tels que [CJP06, MSA05]. Notre problématique principale est de combiner ces différentes facettes. À travers nos différentes contributions nous avons proposé une alternative aux approches actuellement utilisées en RIG. Ces contributions sont :

1. une approche générique d'uniformisation de données afin de les combiner par la suite ;
2. une approche de combinaison de critères étendue par des contraintes permettant à un utilisateur de personnaliser cette combinaison ;
3. un cadre expérimental permettant d'évaluer un SRI géographique ;
4. l'évaluation et la quantification de l'apport de la combinaison des différentes facettes de l'information géographique.

Quatrième partie

Conclusion

Chapitre 9

Conclusion

Sommaire

9.1 Synthèse

Dans le contexte de la recherche d'information géographique, traiter toutes les facettes (spatial, temporel et thématique) pose de nombreux problèmes comme nous l'avons expliqué dans l'état de l'art. Nous avons choisi, comme préconisé dans de nombreux travaux en GIR tels que [CJP06, MSA05], de traiter chacune de ces facettes spécifiquement et de manière indépendante. À cause de cette nécessité de réaliser des traitements dédiés à chaque facette, peu de Systèmes de Recherche d'Information Géographique (SRIG) prennent en charge les trois facettes. Le problème qui se pose ensuite est : comment combiner ces informations hétérogènes ? Plusieurs SRIG ont choisi d'appliquer des approches de filtrage. Néanmoins ce type d'approche ne permet pas d'obtenir un classement des résultats (filtrage parallèle) ou alors uniquement sur un critère (filtrage séquentiel). D'autres ont proposé d'utiliser des approches linéaires standards. Néanmoins, sans homogénéiser les différentes facettes, ce type de combinaison peut introduire des biais, par exemple en avantageant une des facettes. Le problème qui apparait alors est : comment homogénéiser ces différentes informations pour pouvoir les combiner ensuite ?

Dans un premier temps nous avons proposé d'uniformiser les représentations des informations spatiales et temporelles en adoptant la démarche des approches statistiques utilisées en RI sur les termes. Il s'agit, d'une part, d'homogénéiser les formes de représentation de l'information spatiale et de l'information temporelle en une seule, la tuile, et, d'autre part, de mettre en œuvre une même approche statistique de calcul de score

de pertinence. Ce redécoupage en tuiles s'obtient en projetant une représentation à n dimensions sur un tuilage à n dimensions. Dans notre cas une information temporelle est à une dimension, sa représentation est donc projetée sur un tuilage à une dimension (ligne de temps), et, une information spatiale (cartographique) à deux dimensions donc un tuilage à deux dimensions (système de coordonnées planaires). À ces tuiles est associé un poids calculé avec des formules classiques de RI (TF, TF·IDF, OkapiBM25) et le score d'un document résultant d'une requête est calculé via le modèle vectoriel de Salton et le produit scalaire. Les expérimentations ont validé l'intérêt de cette approche d'uniformisation : l'uniformisation spatiale améliore les résultats de RI par rapport à la référence (13 %) et l'uniformisation temporelle donne des résultats de RI équivalents à la référence. Les expérimentations ont aussi montré que le type d'information contenu dans le corpus est prédominant par rapport aux requêtes des utilisateurs et que via une étude statistique automatique il est possible de déterminer le tuilage par défaut à utiliser. **Notre première contribution est donc une méthode d'uniformisation générique que nous avons appliquée à des index de données spatiales et temporelles**.

Une fois ces facettes uniformisées nous avons travaillé sur leur combinaison. Nous avons tout d'abord voulu vérifier l'intérêt de cette combinaison en quantifiant son apport. Nous avons donc testé plusieurs opérateurs de combinaison linéaire reconnus et, avec le meilleur (CombMNZ), nous obtenons 66,4 % d'amélioration qui est significative par rapport aux approches de RI exploitant seulement les termes. Notre expérimentation a montré aussi que ces dimensions ne sont pas redondantes mais bien complémentaires. **Notre deuxième contribution est ainsi l'évaluation et la quantification de l'apport de la combinaison des différentes facettes de l'information géographique**.

Nous avons souhaité améliorer cette combinaison en permettant à un utilisateur de personnaliser une requête multicritère. Dans notre approche, l'utilisateur peut, pour chaque critère, spécifier une exigence (présence obligatoire, optionnelle, ou à exclure), une préférence (critère peu important, important, à éviter), un opérateur (égalité, intersection, ...). Nous avons ainsi identifié différents scénarios de recherche applicables à chaque critère et réalisables grâce à cette approche : exclusion (filtrage négatif), dévalorisation, valorisation, cible, prérequis (filtrage positif). Pour évaluer cette approche, nous l'avons comparée à la meilleure des combinaisons linéaires standards (CombMNZ) et nous obtenons 54,16 % d'amélioration. **Notre troisième contribution propose donc une approche générique de combinaison étendue par contraintes que nous appliquons à la RIG**.

Enfin, pour évaluer notre approche de combinaison PIVcomb, nous avions besoin d'un cadre d'évaluation de SRIG supportant l'ensemble des facettes de l'information géographique. Or les campagnes d'évaluations existantes ont porté sur au maximum deux facettes (spatial et thématique) comme GeoCLEF. **Notre dernière contribution consiste donc en un cadre expérimental permettant d'évaluer un SRIG**. Nous l'avons mis en place en nous appuyant sur les documents fournis par la Médiathèque Intercommunale à Dimension Régionale (MIDR), constituant ainsi une collection de test (MIDR_2010 [41]).

41. La collection de test est accessible sur http://t2i.univ-pau.fr/MIDR/.

9.2 Discussions et Perspectives

9.2.1 Combinaison par contraintes : prise en charge de différents opérateurs

Pour notre approche de combinaison étendue, nous proposons à l'utilisateur de choisir l'opérateur à utiliser pour chaque critère (voir section 6.3). Divers opérateurs sont envisageables : intersection, égalité, inclusion, et proximité. Les opérateurs restent à la charge des SRI utilisés. Notre SRI PIV^2 gère l'intersection et l'égalité (voir section 7.3). Nous allons dans un premier temps détailler les limites actuelles des opérateurs que nous avons mis en œuvre et indiquer comment les autres pourraient être mis en place.

9.2.1.1 Intersection : conflit avec l'exigence d'exclusion et effet de bord dû à l'uniformisation

L'intersection consiste à récupérer les tuiles qui intersectent la représentation de la requête de l'utilisateur (voir section 7.3) et les documents associés. Si l'utilisateur spécifie deux critères, dont un à exclure, il peut y avoir conflit entre les deux critères s'ils s'intersectent. Prenons un exemple : l'utilisateur souhaite trouver les documents évoquant l'année 2001 mais pas ceux de septembre 2001. Si l'opérateur d'intersection est utilisé aucun document évoquant l'année 2001 ne sera retourné étant donné que « septembre 2001 » est inclus dans « année 2001 ».

Une solution est de régler par défaut l'exclusion avec l'opérateur d'égalité car l'exclusion porte généralement sur un critère précis. Sinon il est possible d'alerter l'utilisateur lorsque deux critères entrent en conflit et de lui proposer de modifier sa recherche.

Il faut aussi noter que l'intersection subit un effet de bord dû à l'uniformisation. En effet, si le tuilage utilisé est mensuel et qu'une requête soumise est par exemple « du 1er juin au 15 juin 2000 », la tuile recherchée sera « juin 2000 ». Donc des documents évoquant le 20 juin 2000 ou le 28 juin 2000 pourront être retournés alors qu'ils ne sont pas pertinents. Une solution est d'utiliser une approche de filtrage pour ne conserver que ceux intersectant réellement la requête. Cette approche serait similaire à celle proposée pour l'inclusion (voir section 9.2.1.3).

9.2.1.2 Égalité : problème dû à l'uniformisation

Concernant l'égalité, comme expliqué dans la section 7.3, théoriquement le tuilage ne permet pas de gérer l'égalité d'une information spatiale ou temporelle étant donné que toutes les représentations sont uniformisées dans un tuilage. Nous avons proposé d'utiliser les identifiants des représentations pour filtrer les résultats en ne conservant que ceux contenant la représentation souhaitée. Cette approche est possible car pour chaque triplet (tuile, document, paragraphe), la liste des représentations qui ont été projetées dans la tuile est stockée. Néanmoins, le poids de la tuile reste calculé par rapport à toutes les informations du document. Si nous reprenons l'exemple de « septembre 2001 », le poids de la tuile ne devrait être calculé que par rapport aux évocations de « septembre 2001 » et non pas « l'année 2001 » ou encore « les années 2000 ». Une solution est de compléter

la liste des informations projetées sur la tuile en ajoutant la fréquence correspondante à chaque représentation. Ainsi la fréquence de la tuile ne serait calculée que par rapport à la représentation souhaitée.

9.2.1.3 Ajout d'autres opérateurs : proximité et inclusion

L'opérateur de proximité permet de récupérer les informations proches du critère fournit. Il s'agit donc de modifier la représentation de l'information décrite par ce critère pour inclure ses alentours puis de réaliser une recherche (de type intersection) avec la nouvelle géométrie réalisée. Cet opérateur a été mis en place avec PIV mais n'a pas encore été intégré dans PIV^2.

L'inclusion est plus compliquée à gérer. En effet, tout comme pour l'égalité, le tuilage n'est pas adapté à cet opérateur. Ici le but est de récupérer uniquement les documents dont la représentation est incluse dans celle fournie par l'utilisateur. Néanmoins, cette représentation projetée peut ne pas correspondre à la totalité des tuiles concernées. Prenons l'exemple de la figure 9.1, la requête « entre le 15 mai et le 15 juin 2000 » couvre 50% des tuiles T1 et T2 (mai et juin 2000). L'utilisateur ne souhaite donc pas obtenir des informations sur l'ensemble de ces tuiles. Une solution est, comme pour l'égalité, d'utiliser une technique de filtrage. Ce filtrage consiste à examiner la liste des représentations associées à chaque document résultat et à ne conserver que ceux dans lesquels une des représentations incluses dans celle de la requête (dans la figure 9.1 : et3 et et6) est présente.

9.2.2 De l'importance d'interfaces adaptées

Dans ces travaux nous ne nous sommes pas penchés sur les aspects interface d'interrogation. Pour chaque facette de l'information géographique il est envisagé de proposer une interface d'interrogation et une interface de visualisation adaptées. Par exemple, dans le cadre du prototype PIV une interface Web dédiée à l'information spatiale ainsi qu'une interface Web dédiée à l'information temporelle ont été mises en place. Comme nous pouvons le voir sur la figure 9.2 page 135 pour le spatial et sur la figure 9.4 page 136 pour le temporel, ces interfaces présentent, sous forme graphique, l'interprétation que le système a réalisée à partir d'une requête textuelle. L'utilisateur peut la modifier graphiquement si besoin avant de lancer la recherche. La visualisation des résultats se fait ensuite sur une carte (de type GoogleMaps [42]) pour le spatial (voir figure 9.3 page 135) et sur une ligne de temps pour le temporel (voir figure 9.5 page 136).

Il est désormais nécessaire de réfléchir à la mise en place d'une interface unique permettant de tenir compte des spécificités de chaque facette de l'information géographique tout en restant simple à utiliser. Dans un premier temps, cette interrogation peut se faire via un tableau dans lequel l'utilisateur pourrait choisir les différents critères et les contraintes associées. Dans l'exemple de la figure 9.6 page 137, les exigences et préférences sont illustrées avec des qualificatifs plutôt qu'avec des chiffres car plus faciles à

42. http://maps.google.fr/

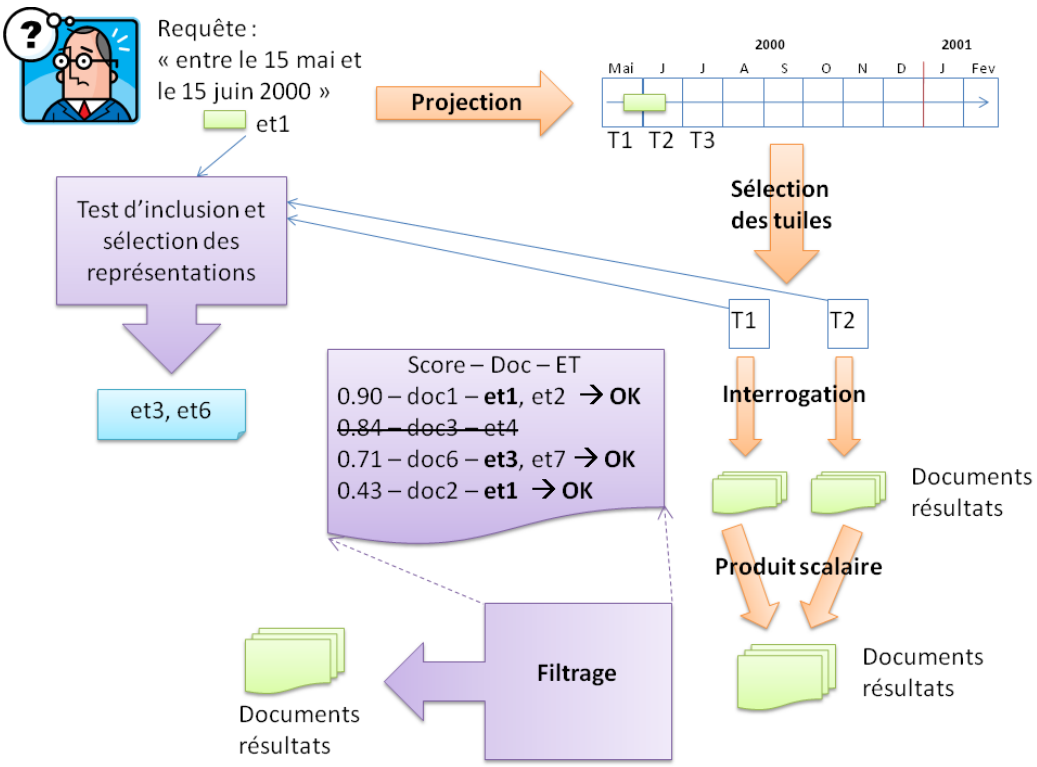


FIGURE 9.1 – Approche possible pour mettre en œuvre l'inclusion

appréhender pour un utilisateur lambda. Il est donc nécessaire de déterminer combien de qualificatifs doivent être proposés et lesquels, mais aussi de vérifier que cette approche est celle qui convient le mieux aux utilisateurs (par exemple, en pratiquant des études d'usages). Une visualisation adaptée au critère peut être intéressante : une carte pour le spatial, une ligne de temps pour le temporel ou une structure hiérarchique de concepts pour le thématique. Si nous reprenons la figure 9.6 page 137, les différentes informations à remplir sont :

- les critères (un critère par colonne) ;
- le type, choix du SRI à utiliser (spatial, temporel ou thématique) ;
- l'exigence, choix parmi : à exclure/optionnel/obligatoire ($-1/0/1$ dans PIVcomb) ;
- la préférence, choix par exemple parmi : à éviter/inintéressant/indifférent/peu intéressant/intéressant/très intéressant (par exemple $-1/-0.5/0/0.3/0.6/1$ dans PIVcomb) ;
- l'opérateur, choix parmi : intersection/égalité/inclusion/proximité ;
- l'interprétation, retournée par le SRI (une fois la requête soumise), afin que l'utilisateur puisse vérifier que cela corresponde à ce qu'il cherche et puisse modifier si ce n'est pas le cas.

Dans un deuxième temps, il est envisageable de proposer à l'utilisateur, comme actuellement dans PIV, de soumettre sa requête en texte libre (par exemple « montagnes dans les Pyrénées entre 1800 et 1900 à l'exception de Gavarnie et si possible sans rapport avec les ascensions ») tel que dans l'exemple de la figure 9.7 page 137. Puis, une fois que le système a analysé la requête, il pourrait générer le tableau de la figure 9.6 page 137.

Pour l'affichage des résultats, il est envisageable de les présenter sous forme de liste comme Google mais aussi d'enrichir ces résultats en précisant, pour chaque document, le score obtenu pour chaque critère. Il est aussi possible d'afficher ces résultats graphiquement, par exemple sur une carte et une ligne de temps.

9.2.3 Autres perspectives

D'autres propositions de perspectives nécessitent plus de réflexion :

- **Amélioration du traitement de la facette thématique** : pour l'instant, comme le font les SRIG actuels, nous utilisons la facette thématique uniquement de manière limitée à l'usage des termes extraits du document. Néanmoins, fréquemment, plusieurs termes portent sur le même thème (exemple : automobile et voiture). Or via une approche par termes, ils ne peuvent pas être reliés. Utiliser des ressources supplémentaires (telles que des ontologies) pour extraire des thèmes plutôt que des termes permettrait encore d'améliorer la recherche d'information. Nous pourrions ensuite appliquer notre approche d'uniformisation sur les thèmes pour les pondérer selon leur fréquence d'apparition.
- **Mise en place de la pondération minimale pour le calcul de fréquence proportionnelle** : pour éviter de trop pénaliser les représentations ponctuelles, nous avons proposé de mettre en place une pondération minimale. Néanmoins, des expérimentations doivent être menées afin de déterminer la valeur minimale optimale.
- **Appliquer l'approche de combinaison par contraintes à d'autres critères** : l'approche que nous proposons n'est pas limitée à l'information géographique. Elle peut être appliquée à d'autres informations et donc il serait intéressant de tester son apport dans ces cas là. Par exemple, cette approche peut être appliquée à un SRI classique. Chaque critère serait traité par PIVcomb comme autant d'appels distincts au SRI. Ce type original d'encapsulation d'un SRI existant permet notamment de l'enrichir d'un langage d'interrogation permettant d'exprimer des contraintes d'exigence et de préférence.

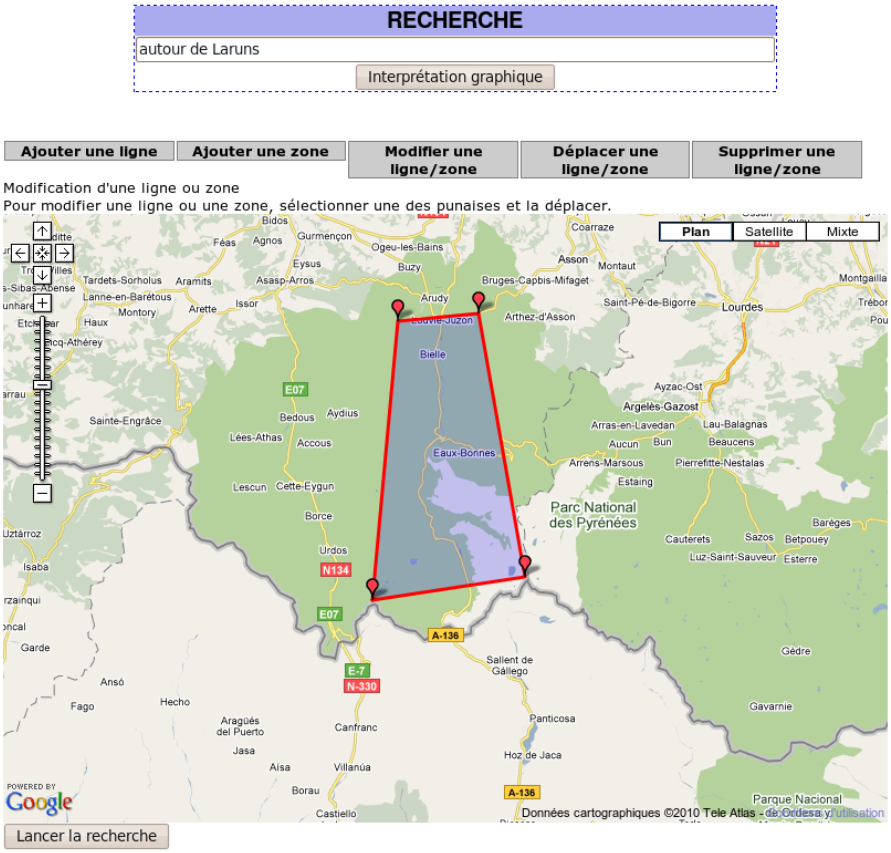


FIGURE 9.2 – Interface d'interrogation spatiale : interprétation de la requête

FIGURE 9.3 – Interface d'interrogation spatiale : affichage des résultats

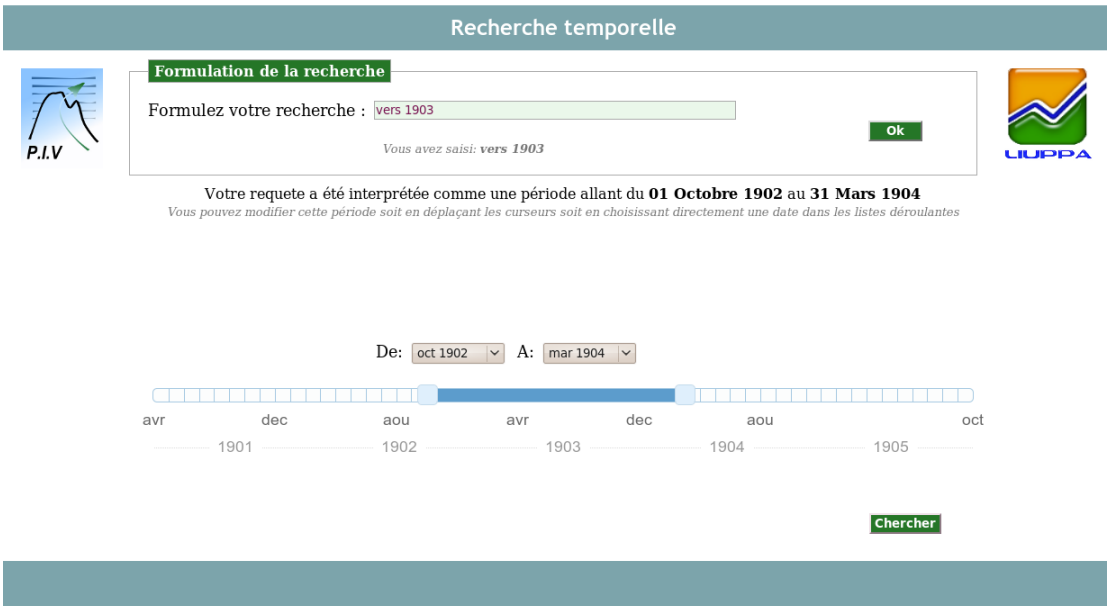


FIGURE 9.4 – Interface d'interrogation temporelle : interprétation de la requête

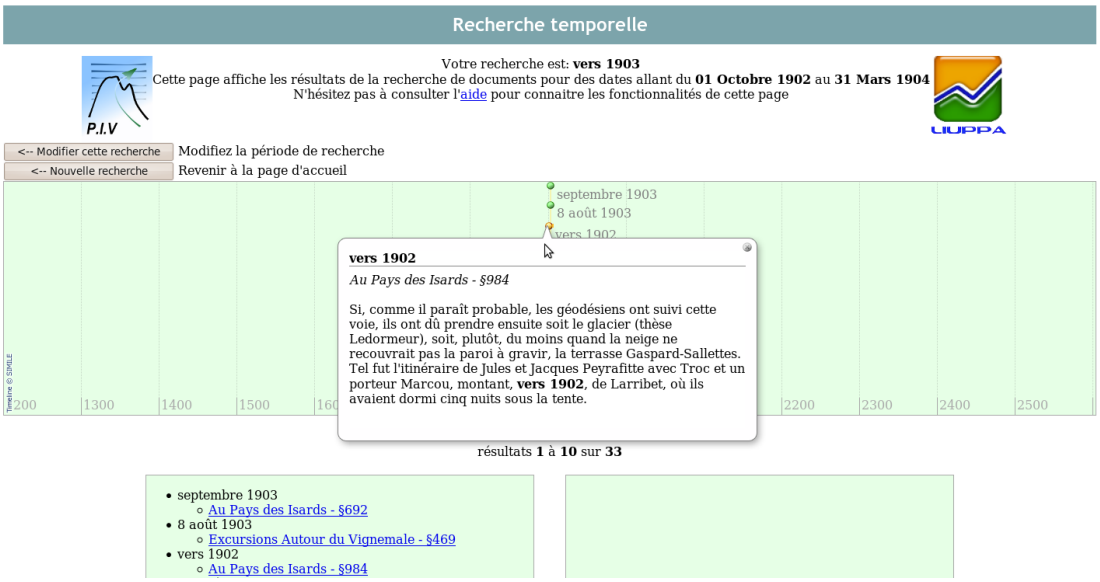


FIGURE 9.5 – Interface d'interrogation temporelle : affichage des résultats

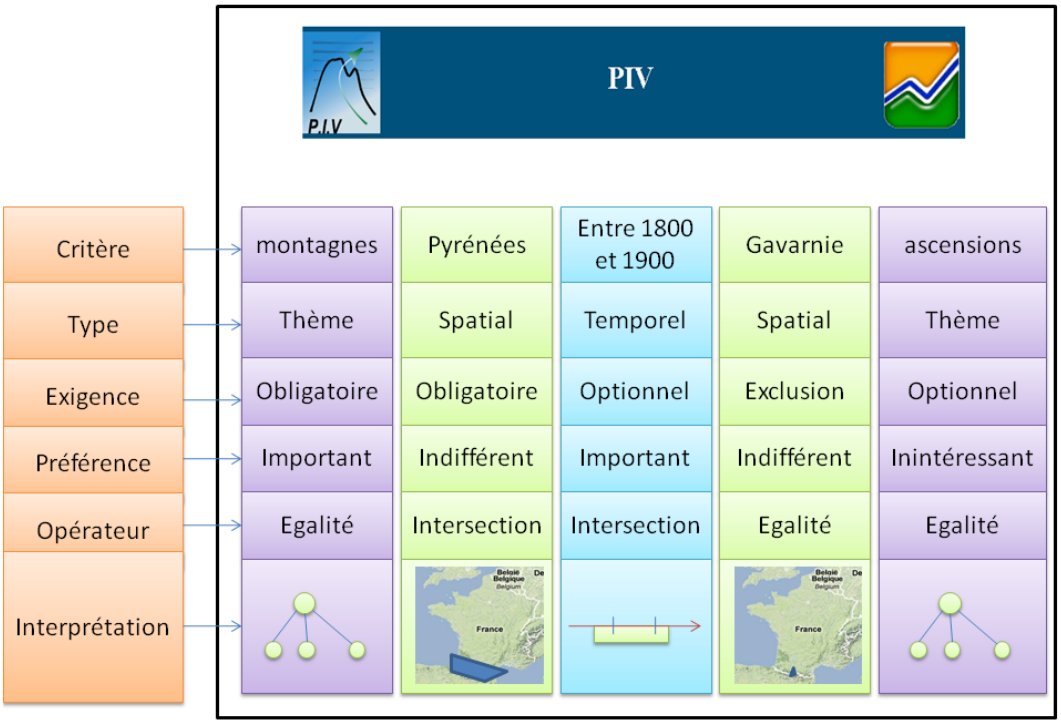


FIGURE 9.6 – Exemple d'interface illustrant l'interprétation de la requête par le système et permettant de corriger si besoin est

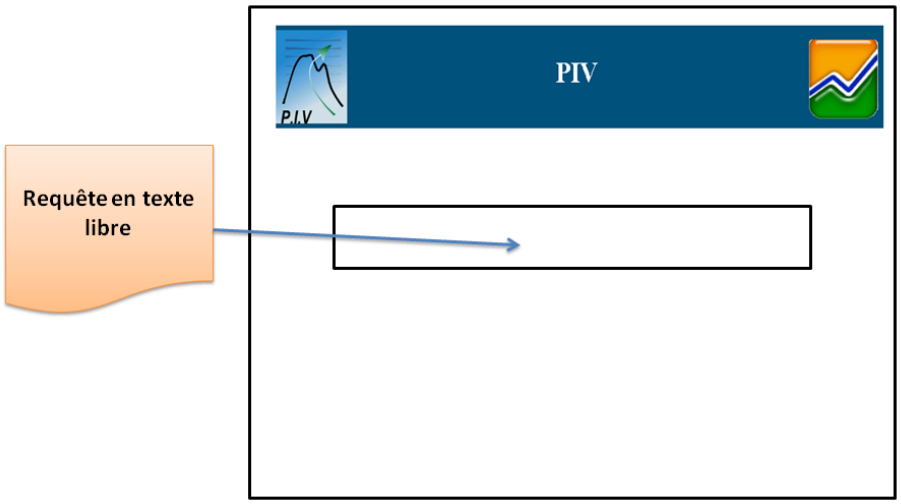


FIGURE 9.7 – Exemple d'interface d'interrogation simple

Bibliographie

[AGQ06] Stéphane Ayache, Jérôme Gensel, and Georges Quénot. CLIPS-LSR Experiments at TRECVID 2006. In *TREC Workshop on Video Retrieval Evaluation*, 2006.

[All84] James Allen. Towards a General Theory of Action and Time. *Artificial Intelligence*, 23(2):123–154, 1984.

[And10] Geoffrey Andogah. *Geographically Constrained Information Retrieval*. PhD thesis, University of Groningen, Netherlands, May 2010.

[BB99] Jon Bosak and Tim Bray. XML and the Second-Generation Web. *Scientific American*, 280(5):89–93, May 1999.

[BCEM03] Frédérik Bilhaut, Thierry Charnois, Patrice Enjalbert, and Yann Mathet. Geographic reference analysis for geographic document querying. In *HLT-NAACL'03: Proceedings of the workshop on Analysis of geographic references*, pages 55–62, Morristown, NJ, USA, 2003. ACL.

[BCJ+05] Bénédicte Bucher, Paul Clough, Hideo Joho, Ross Purves, and Awase Khirni Syed. Geographic IR Systems: Requirements and Evaluation. In *ICC'05: Proceedings of the 22nd International Cartographic Conference*. Global Congressos, 2005. CDROM.

[BDEH07] Frédérik Bilhaut, Franck Dumoncel, Patrice Enjalbert, and Nicolas Hernandez. Indexation sémantique et recherche d'information interactive. In *CORIA'07: Actes de la 4^e Conférence en Recherche d'Information et Applications*, pages 65–76. Université de Saint-Étienne, 2007.

[Bes04] Romaric Besançon. *Technologies Statistiques pour la recherche d'informations : les modèles vectoriels*, volume Les systèmes de recherche d'informations, chapter 2, pages 35–54. Hermès-Lavoisier, 2004.

[BGL07] Michael K. Buckland, Fredric C. Gey, and Ray R. Larson. Access to Heritage Resources Using What, Where, When, and Who. In *Museums and the Web 2007: Proceedings*, 2007.

[BL04] Michael Buckland and Lewis Lancaster. Combining Place, Time, and Topic: The Electronic Cultural Atlas Initiative. *D-Lib Magazine*, 10, May 2004.

[BPP+93] Denis Bouyssou, Patrice Perny, Marc Pirlot, Alexis Tsoukiàs, and Philippe Vincke. A Manifesto for the New MCDM Era. *Journal of Multi-Criteria Decision Analysis*, 2:125–127, 1993.

[BR10] Gülçin Büyüközkan and Da Ruan. Choquet integral based aggregation approach to software development risk assessment. *Information Sciences*, 180(3):441 – 451, 2010.

[BV00] Chris Buckley and Ellen M. Voorhees. Evaluating Evaluation Measure Stability. In *SIGIR'00: Proceedings of the 23rd international ACM SIGIR conference*, pages 33–40, New York, NY, USA, 2000. ACM.

[BW06] Frédérik Bilhaut and Antoine Widlöcher. Linguastream: an integrated environment for computational linguistics experimentation. In *EACL'06: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pages 95–98, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[BYRN99] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, 1999.

[Cai02] Guoray Cai. GeoVSM: An Integrated Retrieval Model for Geographic Information. In Max J. Egenhofer and David M. Mark, editor, *GIScience*, volume 2478 of *Lecture Notes in Computer Science*, pages 65–79. Springer, 2002.

[CDG+07] Yih-Farn Chen, Giuseppe Di Fabbrizio, David Gibbon, Rittwik Jana, Serban Jora, Bernard Renger, and Bin Wei. GeoTracker: Geospatial and Temporal RSS Navigation. In Carey L. Williamson and Mary Ellen Zurko and Peter F. Patel-Schneider and Prashant J. Shenoy, editor, *Proceedings of the 16th International World Wide Web Conference*, pages 41–50, Banff, Alberta, May 2007. ACM Press.

[CJP06] Paul Clough, Hideo Joho, and Ross Purves. Judging the Spatial Relevance of Documents for GIR. In *ECIR'06: Proceedings of the 28th European Conference on IR Research*, volume 3936 of *Lecture Notes in Computer Science*, pages 548–552. Springer, 2006.

[CMBT02] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: an architecture for development of robust HLT applications. In *ACL'02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 168–175, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[CMS09] Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA, 2009.

[CPDP09a] Célia Costa Pereira, Mauro Dragoni, and Gabriella Pasi. Multidimensional relevance: A new aggregation criterion. In *ECIR'09: Proceedings*

of the 31th European Conference on IR Research on Advances in Information Retrieval, pages 264–275, Berlin, Heidelberg, 2009. Springer-Verlag.

[CPDP09b] Célia Costa Pereira, Mauro Dragoni, and Gabriella Pasi. A prioritized "and" aggregation operator for multidimensional relevance assessment. In Roberto Serra and Rita Cucchiara, editors, *AI*IA*, volume 5883 of *Lecture Notes in Computer Science*, pages 72–81. Springer, 2009.

[EI05] Garbinand Eric and Maniand Inderjeet. Disambiguating toponyms in news. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 363–370, Vancouverand British Columbiaand Canada, October 2005. Association for Computational Linguistics.

[EMH94] Max J. Egenhofer, David M. Mark, and John Herring. The 9-Intersection: Formalism and Its Use for Natural- Language Spatial Predicates. Technical report, NCGIA, 1994.

[Flo09] Luciano Floridi. Semantic Conceptions of Information. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab Center for the Study of Language and Information Stanford University Stanford, CA 94305-4115, summer 2009 edition, 2009.

[Flu04] Christian Fluhr. *L'évaluation des systèmes de recherche d'informations textuelles*, volume Evaluation des systèmes de traitement de l'information, chapter 1, pages 27–44. Hermes-Lavoisier, 2004.

[FS93] Edward A. Fox and Joseph A. Shaw. Combination of Multiple Searches. In Donna K. Harman, editor, *TREC-1: Proceedings of the First Text REtrieval Conference*, pages 243–252, Gaithersburg, MD, USA, February 1993. NIST.

[Gai01] Mauro Gaio. Traitements de l'Information Géographique : Représentations et Structures. In *Mémoire d'HDR, Université de Caen*, 2001.

[GAMS08] Qingqing Gan, Josh Attenberg, Alexander Markowetz, and Torsten Suel. Analysis of geographic queries in a search engine log. In *LocWeb'08: Proceedings of the first international workshop on Location and the web*, pages 49–56, New York, NY, USA, 2008. ACM.

[GD09] Ayse Goker and John Davies. *Information retrieval: Searching in the 21st Century*. John Wiley & Sons, 2009.

[GGHR00] Eric Gaussier, Gregory Grefenstette, David Hull, and Claude Roux. Recherche d'information en français et traitement automatique des langues. *Traitement automatique des langues*, 41(2):473–493, 2000.

[GH05] Otis Gospodnetić and Erik Hatcher. *Lucene in Action*. Manning Publications, 2005.

[GLK+10] Fredric Gey, Ray Larson, Noriko Kando, Jorge Machado, and Tetsuya Sakai. NTCIR-GeoTime Overview: Evaluating Geographic and Temporal Search. In *NTCIR'10: Proceedings of the 8th NTCIR Workshop*, pages 147–153, Tockyo, Japan, 2010. NII.

[GLS+05] Fredric C. Gey, Ray R. Larson, Mark Sanderson, Hideo Joho, Paul Clough, and Vivien Petras. GeoCLEF'05: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In *CLEF'05: Proceedings of the 6th workshop on Cross-Language Evalution Forum*, volume 4022 of *LNCS*, pages 908–919. Springer, 2005.

[GSE+08] Mauro Gaio, Christian Sallaberry, Patrick Etcheverry, Christophe Marquesuzaa, and Julien Lesbegueries. A global process to access documents' contents from a geographical point of view. In *Journal of Visual Languages And Computing*, volume 19, pages 3–23, Orlando, FL, USA, 2008. Academic Press, Inc.

[Har05] Donna K. Harman. The TREC Test Collections. In Ellen M. Voorhees and Donna K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 2, pages 21–53. MIT Press, 2005.

[HL09] Djamila Hamdadou and Thérèse Libourel. Couplage approche multicritère et négociation pour l'aide à la décision en aménagement du territoire. In *SAGEO 2009*, 2009.

[Hul93] David Hull. Using Statistical Testing in the Evaluation of Retrieval Experiments. In *SIGIR'93: Proceedings of the 16th annual international ACM SIGIR conference*, pages 329–338, New York, NY, USA, 1993. ACM Press.

[JK02] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[JP08] Christopher B. Jones and Ross S. Purves. Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3):219–228, 2008.

[JZR+08] Rosie Jones, Wei V. Zhang, Benjamin Rey, Pradhuman Jhala, and Eugene Stipp. Geographic intention and modification in web search. *International Journal of Geographical Information Science*, 22(3):229–246, 2008.

[Kow97] Gerald Kowalski. *Information Retrieval Systems : Theory and Implementation*. Kluwer Academic Publishers, 1997.

[Lar96] Ray R. Larson. Geographic information retrieval and spatial browsing. In Smith and M. Gluck, editors, *Geographic Information Systems and Libraries: Patronsand Mapsand and Spatial Information*, pages 81–124, 1996.

[Lee97] Joon Ho Lee. Analyses of Multiple Evidence Combination. In *SIGIR'97: Proceedings of the 20th annual international ACM SIGIR conference*, pages 267–276, New York, NY, USA, 1997. ACM Press.

[Les07] Julien Lesbegueries. *Plate-forme pour l'indexation spatiale multi-niveaux d'un corpus territorialisé*. PhD thesis, Université de Pau et des Pays de l'Adour, 2007.

[LF73] Frederick Wilfrid Lancaster and Emily Gallup Fayen. *Information retrieval: on-line*. Melville Pub. Co. Los Angeles, 1973.

[LF04] Ray R. Larson and Patricia Frontiera. Spatial Ranking Methods for Geographic Information Retrieval (GIR) in Digital Libraries. In Rachel Heery and Liz Lyon, editor, *ECDL*, volume 3232 of *Lecture Notes in Computer Science*, pages 45–56. Springer, 2004.

[LG03] Christophe Labreuche and Michel Grabisch. The Choquet integral for the aggregation of interval scales in multicriteria decision making. *Fuzzy Sets and Systems*, 137(1):11 – 26, 2003.

[LG08] Leila Liberge and Job Gerlings. Cultural Heritage on the (Geographical) Map. In *Museums and the Web 2008: Proceedings*, pages 163–172, 2008.

[LGL06] Julien Lesbegueries, Mauro Gaio, and Pierre Loustau. Geographical information access for non-structured data. In *SAC'06: Proceedings of the 2006 ACM symposium on Applied computing*, pages 83–89, New York, NY, USA, 2006. ACM.

[LGMR05] Paul A. Longley, Michael F. Goodchild, David J. Maguire, and David W. Rhind. *Geographic Information Systems and Science*. John Wiley & Sons, 2005.

[LGS07] Annig Le Parc-Lacayrelle, Mauro Gaio, and Christian Sallaberry. La composante temps dans l'information géographique textuelle. *Revue Document Numérique*, 10(2):129–148, 2007.

[LMSC06] Yi Li, Alistair Moffat, Nicola Stokes, and Lawrence Cavedon. Exploring Probabilistic Toponym Resolution for Geographical Information Retrieval. In Ross Purves and Chris Jones, editor, *GIR*, pages 17–22. Department of Geography, University of Zurich, 2006.

[Lou08] Pierre Loustau. *Interprétation automatique d'itinéraires dans des récits de voyages*. PhD thesis, Université de Pau et des Pays de l'Adour, 2008.

[LSDJ06] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):1–19, 2006.

[LSSS07] Michael D. Lieberman, Hanan Samet, Jagan Sankaranarayanan, and Jon Sperling. STEWARD: architecture of a spatio-textual search engine. In *GIS '07: Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*, pages 1–8, New York, NY, USA, 2007. ACM.

[Man03] Inderjeet Mani. Recent developments in temporal information extraction. In Nicolas Nicolov and Kalina Bontcheva and Galia Angelova and Ruslan Mitkov, editor, *RANLP*, volume 260 of *Current Issues in Linguistic Theory (CILT)*, pages 45–60. John Benjamins, Amsterdam/Philadelphia, 2003.

[Mar99] Jean-Marc Martel. L'aide multicritère à la décision : méthodes et applications. In *Actes de la conférence CORS - SCRO*, pages 6–16, 1999.

[MBP+07] Bruno Martins, José Luis Borbinha, Gilberto Pedrosa, João Gil, and Nuno Freire. Geographically-aware information retrieval for collections of digitized historical maps. In Ross Purves and Chris Jones, editor, *GIR*, pages 39–42. ACM, 2007.

[MCF+03] Jacek Malczewski, Terry Chapman, Cindy Flegel, Dan Walters, Dan Shrubsole, and Martin A Healy. GIS - multicriteria evaluation with ordered weighted averaging (OWA): case study of developing watershed management strategies. *Environment and Planning A*, 35(10):1769–1784, October 2003.

[Mei87] Alphonse Meillon. *Excursions autour du Vignemale dans les hautes vallées de Cauterets, de Gavarnie et du Rio Aran en Aragon.* Sirius, 1987.

[MHR+08] Inderjeet Mani, Janet Hitzeman, Justin Richer, Dave Harris, Rob Quimby, and Ben Wellner. SpatialML: Annotation Scheme, Corpora, and Tools. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008.

[MMB08] Bruno Martins, Hugo Manguinhas, and José Luis Borbinha. Extracting and Exploring the Geo-Temporal Semantics of Textual Resources. In *ICSC*, pages 1–9. IEEE Computer Society, 2008.

[MMBS09] Hugo Manguinhas, Bruno Martins, José Luis Borbinha, and Willington Siabato. The DIGMAP geo-temporal Web gazetteer service. In *e-Perimetron: International Web journal on sciences and technologies affined to history of cartography and maps*, volume 4, pages 9–24, 2009.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval.* Cambridge University Press, New York, 2008.

[MSA05] Bruno Martins, Mário J. Silva, and Leonardo Andrade. Indexing and ranking in Geo-IR systems. In *GIR '05: Proceedings of the 2005 workshop on Geographic information retrieval*, pages 31–34, New York, NY, USA, 2005. ACM.

[MSZ03] Inderjeet Mani, Barry Schiffman, and Jianping Zhang. Inferring Temporal Ordering of Events in News. In *HLT-NAACL*, pages 55–57, 2003.

[MT04] Philippe Muller and Xavier Tannier. Annotating and measuring temporal relations in texts. In *Proceedings of the 20th International Conference on Computational Linguistics (Coling 04)*, pages 50–56, Genève, Suisse, aug 2004. Association for Computational Linguistics.

[NRD08] Sérgio Nunes, Cristina Ribeiro, and Gabriel David. Use of Temporal Expressions in Web Searches. In *ECIR'08: Proceedings of the 30th European Conference on Information Retrieval*, pages 580–584, 2008.

[OAP+05] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. Terrier Information Retrieval Platform. In *ECIR'05: Proceedings of the 27th European Conference on IR Research*, volume 3408 of *LNCS*, pages 517–519. Springer, 2005.

[OC01] Paul Ogilvie and James P. Callan. Experiments Using the Lemur Toolkit. In *TREC'01: Proceedings of the 9th Text REtrieval Conference*, pages 103–108, Gaithersburg, MD, USA, 2001. NIST.

[Paz97] Maria Teresa Pazienza. *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, volume 1299. Springer, Berlin Heidelberg New York, 1997.

[Paz99] Maria Teresa Pazienza. *Information Extraction: Towards Scalable, Adaptable Systems*, volume 1714 of *Lecture Notes in Computer Science*. Springer, 1999.

[PCI+03] James Pustejovsky, José M. Castaño, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. TimeML: Robust Specification of Event and Temporal Expressions in Text. In Mark T. Maybury, editor, *New Directions in Question Answering*, pages 28–34. AAAI Press, 2003.

[PCSH10a] Damien Palacio, Guillaume Cabanac, Christian Sallaberry, and Gilles Hubert. Cadre d'évaluation de SRI géographique : apport de la combinaison des dimensions spatiale, temporelle et thématique. In *INFORSID'10 : 28e congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision*, pages 245–260. Éditions Inforsid, May 2010.

[PCSH10b] Damien Palacio, Guillaume Cabanac, Christian Sallaberry, and Gilles Hubert. Measuring Effectiveness of Geographic IR Systems in Digital Libraries: Evaluation Framework and Case Study. In Mounia Lalmas, Joemon Jose, Andreas Rauber, Fabrizio Sebastiani, and Ingo Frommholz, editors, *ECDL'10: Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries*, volume 6273 of *LNCS*, pages 340–351. Springer, September 2010.

[PEH+09] Dieter Pfoser, Alexandros Efentakis, Thanasis Hadzilacos, Sophia Karagiorgou, and Giorgos Vasiliou. Providing Universal Access to History Textbooks: A Modified GIS Case. In James D. Carswell, A. Stewart

Fotheringham, and Gavin McArdle, editors, *W2GIS*, volume 5886 of *Lecture Notes in Computer Science*, pages 87–102. Springer, 2009.

[Pet01] Carol Peters. Introduction. In *CLEF'01: Proceedings of the 1st Workshop Cross-Language Information Retrieval and Evaluation*, volume 2069 of *LNCS*, pages 1–6. Springer, 2001.

[PKLS05] James Pustejovsky, Robert Knippen, Jessica Littman, and Roser Sauri. Temporal and Event Information in Natural Language Text. *Language Resources and Evaluation*, 39:123–164, 2005.

[PMLC07] Trong-Ton Pham, Nicolas Maillot, Joo-Hwee Lim, and Jean-Pierre Chevallet. Latent semantic fusion model for image retrieval and annotation. In Mário J. Silva and Alberto H. F. Laender and Ricardo A. Baeza-Yates and Deborah L. McGuinness and Bjørn Olstad and Øystein Haug Olsen and André O. Falcão, editor, *CIKM*, pages 439–444. ACM, 2007.

[POGCGVUL08] José M. Perea-Ortega, Miguel A. García-Cumbreras, Manuel García-Vega, and L. A. Ureña-López. Comparing Several Textual Information Retrieval Systems for the Geographical Information Retrieval Task. In *NLDB'08: Proceedings of the 13th international conference on Natural Language and Information Systems*, pages 142–147, Berlin, Heidelberg, 2008. Springer-Verlag.

[Poi03] Thierry Poibeau. *Extraction automatique d'information.* Hermès-Lavoisier, 2003.

[Pow97] Daniel J. Power. What is a DSS? *The On-Line executive Journal for Data-Intensive Decision Support*, 1(3), 1997.

[PSG09] Damien Palacio, Christian Sallaberry, and Mauro Gaio. Normalizing Spatial Information to Better Combine Criteria in Geographical Information Retrieval. In *ECIR-GIIW'09: Proceeding of the international workshop on Geographic Information on the Internet*, pages 37–48, 2009.

[PSG10] Damien Palacio, Christian Sallaberry, and Mauro Gaio. Normalizing Spatial Information to Improve Geographical Information Indexing and Retrieval in Digital Libraries. In *ISGIS'10: Proceedings of the Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science proceedings*, pages 229–234, 2010.

[RCC92] David A. Randell, Zhan Cui, and Anthony Cohn. A spatial logic based on regions and connection. In Bernhard Nebel, Charles Rich, and William Swartout, editors, *KR'92. Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Conference*, pages 165–176. Morgan Kaufmann, San Mateo, California, 1992.

[RPR99] Martin Riegel, Jean-Chirstophe Pellat, and René Rioul. *Grammaire méthodique du français*. Presses Universitaires de France (PUF), 1999.

[Saa80] Thomas L. Saaty. *The Analytic Hierarchy Process, Planning, Priority Setting, Resource Allocation*. McGraw-Hill, New york, 1980.

[Sal68] Gerard Salton. *Automatic Information Organization and Retrieval*. McGraw Hill Text, 1968.

[Sal71] Gerard Salton. *The SMART retrieval system: Experiments in automatic document processing*. Prentice-Hall, Upper Saddle River, NJ, 1971.

[Sam84] Hanan Samet. The Quadtree and Related Hierarchical Data Structures. *ACM Comput. Surv.*, 16(2):187–260, 1984.

[San10] Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

[SBLG07] Christian Sallaberry, Mustapha Baziz, Julien Lesbegueries, and Mauro Gaio. Towards an IE and IR System Dealing with Spatial Information in Digital Libraries – Evaluation Case Study. In *ICEIS'07: Proceedings of the 9th International Conference on Enterprise Information Systems*, pages 190–197, 2007.

[SC10] Diana Santos and Luís Cabral. GikiCLEF: Expectations and Lessons Learned. In Carol Peters, Giorgio Di Nunzio, Mikko Kurimo, Djamel Mostefa, Anselmo Penas, and Giovanna Roda, editors, *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, volume 6241 of *LNCS*, pages 212–222. Springer Berlin / Heidelberg, 2010.

[Sch94] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK, 1994.

[SD08] Jacques Savoy and Ljiljana Dolamic. Variations autour de tf idf et du moteur Lucene. In *9es Journées internationales d'Analyse statistique des Données Textuelles*, pages 1047–1058, 2008.

[SGPL08] Christian Sallaberry, Mauro Gaio, Damien Palacio, and Julien Lesbegueries. Fuzzying GIS Topological Functions for GIR Needs. In *GIR'08: Proceeding of the 2nd international workshop on Geographic information retrieval*, pages 1–8, New York, NY, USA, 2008. ACM.

[SJvR75] Karen Spärck Jones and Cornelis Joost van Rijsbergen. Report on the need for and provision of an 'ideal' information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

[SK04] Mark Sanderson and Janet Kohler. Analyzing Geographic Queries. In *Proceedings of the Workshop on Geographic Information Retrieval, SIGIR*, 2004.

[SM83] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[SYY75] Gerard Salton, Chung-Shu Yang, and Clement T. Yu. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1):33–44, 1975.

[SZ05] Mark Sanderson and Justin Zobel. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In *SIGIR'05: Proceedings of the 28th annual international ACM SIGIR conference*, pages 162–169, New York, NY, USA, 2005. ACM.

[TM08] Xavier Tannier and Philippe Muller. Evaluation Metrics for Automatic Temporal Annotation of Texts. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008.

[TWFM07] Ye Tian, Gary Weiss, D. Frank, and Hsu Qiang Ma. A Combinatorial Fusion Method for Feature Mining. In *Proceedings of KDD'07 Workshop on Mining Multiple Information Sources*, 2007.

[Use96] E. Lynn Usery. A feature-based geographic information system model. *Photogrammetric Engineering & Remote Sensing*, 62(7):833–838, 1996.

[Val06] Valcartier. GRID - Geospatial Retrieval of Indexed Document. Technical report, R&D pour la défense Canada, 2006.

[VC99] Christopher C. Vogt and Garrison W. Cottrell. Fusion Via a Linear Combination of Scores. *Inf. Retr.*, 1(3):151–173, 1999.

[VGS+09] Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. The TempEval challenge: identifying temporal relations in text. *Lang. Resour. Eval.*, 43(2):161–179, 2009.

[VH05] Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, MA, USA, 2005.

[VJJS05] Subodh Vaid, Christopher B. Jones, Hideo Joho, and Mark Sanderson. Spatio-textual Indexing for Geographical Search on the Web. In Claudia Bauzer Medeiros and Max J. Egenhofer and Elisa Bertino, editor, *SSTD*, volume 3633 of *Lecture Notes in Computer Science*, pages 218–235. Springer, 2005.

[VKRAVZ05] Marc Van Kreveld, Iris Reinbacher, Avi Arampatzis, and Roelof Van Zwol. Multi-Dimensional Scattered Ranking Methods for Geographic Information Retrieval. In *GeoInformatica*, volume 9, pages 61–84, Hingham, MA, USA, 2005. Kluwer Academic Publishers.

[Voo99] Ellen M. Voorhees. Natural language processing and information retrieval. In Maria Teresa Pazienza, editor, *SCIE*, volume 1714 of *Lecture Notes in Computer Science*, pages 32–48. Springer, 1999.

[Voo01] Ellen M. Voorhees. Evaluation by Highly Relevant Documents. In *SIGIR'01: Proceedings of the 24th annual international ACM SIGIR conference*, pages 74–82, New York, NY, USA, 2001. ACM.

[Voo07] Ellen M. Voorhees. TREC: Continuing information retrieval's tradition of experimentation. *Commun. ACM*, 50(11):51–54, 2007.

[WB05] Antoine Widlöcher and Frédérik Bilhaut. La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus. In *TALN'05: Actes de la 12^e Conférence sur le Traitement Automatique du Langage Naturel*, pages 517–522, 2005.

[Wol92] David H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[Yag88] Ronald R. Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Trans. Syst. Man Cybern.*, 18(1):183–190, 1988.

[Zel82] Milan Zeleny. *Multiple Criteria Decision Making*. McGraw-Hill, New York, 1982.

Résumé

Des études récentes montrent une part croissante de requêtes sur les moteurs de recherche du Web comportant des critères géographiques. Cette part est encore plus conséquente sur des corpus plus spécifiques tels que des documents patrimoniaux (récits de voyages par exemple). On admet que l'information géographique est composée de trois facettes : le spatial, le temporel et le thématique. Les travaux effectués dans notre laboratoire visent l'extraction et la construction d'index indépendants et spécifiques aux trois facettes (spatiales, temporelles et thématiques). L'objet de ce travail de thèse est de combiner les trois facettes pour effectuer des recherches multicritère.

Ce travail s'intègre au croisement de plusieurs disciplines : Traitement Automatique des Langages Naturels (TALN), Systèmes d'Information Géographique (SIG), Recherche d'Information classique (RI) et Recherche d'Information Géographique (RIG).

Notre première contribution porte sur une méthode originale de combinaison des index spécifiques. Lors de l'interrogation il s'agit de questionner de manières indépendantes les différents index puis de combiner les listes de résultats restitués lors de leur interrogation. De plus, nous proposons à un utilisateur de personnaliser cette combinaison par des contraintes. Pour pouvoir effectuer cette combinaison, nous proposons d'imiter les approches d'homogénéisation utilisées dans les stratégies de RI classiques portant sur des termes et les lemmes correspondants. Pour les informations géographiques il s'agit de les redécouper en tuiles et de travailler sur leur fréquence d'apparition. Notre deuxième contribution porte sur une approche d'uniformisation générique mise en œuvre sur l'information spatiale et l'information temporelle. Afin d'évaluer ces différentes propositions, nous les avons testées et validées via différents prototypes et expérimentations. La dernière contribution consiste en un cadre d'évaluation d'un système de recherche géographique. Grâce à ce cadre nous avons pu vérifier et quantifier l'apport de la combinaison de critères géographiques ainsi que comparer différentes approches de combinaisons.

Mots-clés: Recherche d'information Géographique, Approche générique d'uniformisation de l'information, Combinaison de résultats, Combinaison par contraintes, Cadre d'évaluation de systèmes de RI géographique

Abstract

Recent studies show an increasing proportion of queries with geographic criteria on Web search engines. This part is even bigger on specific corpora like cultural heritage collection (e.g. travelogues). We admit that the geographic information is composed of three facets: spatial, temporal and thematic. Works realized in our laboratory aim geographic information extraction from textual documents and the construction of independent and specific indexes for theses three facets. The goal of this thesis is to combine these three facets to support multicriteria searches.

This work concerns several fields: Natural Language Processing (NLP), Geographic Information System (GIS), classic Information Retrieval (IR) and Geographic Information Retrieval (GIR).

Our first contribution is about an original combination approach of specific indexes. During the retrieval process, it consists first in querying the different indexes independently and then combining the results lists. We propose also a user to personalize this combination with constraints. In order to realize this combination, we propose to imitate the homogenization approaches used in classical IR strategies that represent terms with corresponding lemmas. For geographic information, it consists in segmenting them on tiles and on using their occurrence frequency. So, our second contribution concerns a generic standardization approach implemented on spatial and temporal information. In order to evaluate these different propositions, we have tested and validated them via several prototypes and experimentations. The last contribution relates to an evaluation framework for GIR systems. Thanks to this framework, we verified and quantified the benefit of combining the different geographic information facets and also have compared several combination approaches.

Keywords: Geographic Information Retrieval, Generic information standardization approach, Results combination, Constraints combination, Evaluation framework for GIRS