Introduction

I.1. Recherche d'information géographique

Les travaux présentés dans cet ouvrage s'inscrivent dans le domaine de la recherche d'information géographique (RIG). La recherche d'information (RI) est la tâche de recherche, au sein d'une collection de documents généralement stockée sur Internet, de documents satisfaisant un besoin d'information [MAN 08b]. La RIG, nommée et définie pour la première fois par Ray Larson [LAR 96], est la tâche de recherche de documents vérifiant des caractéristiques géographiques : ainsi, les zones géographiques évoquées dans les documents résultant d'une RIG recouvrent partiellement ou totalement celles exprimées dans la requête. La série de conférences GIR[1] débutée en 2004 [PUR 04] a fortement contribué au développement de la RIG. La RIG est focalisée sur la dimension spatiale dans un premier temps puis, pour les documents textuels, étendue par la dimension thématique véhiculée par des termes porteurs de sens (autre que spatial). Nous retrouvons la dimension *spatio-textual search* dans la série de conférences SSTD[2] [JEN 01] à partir de 2005 [VAI 05], GeoCLEF[3] [GEY 05] à partir de 2005 [BUC 05] et GIS[4] [PIS 93] à partir de 2007 [LIE 07]. Ce sont

1. www.geo.uzh.ch/ rsp/gir10/.
2. http://dblab.cs.ucr.edu/conferences/sstd01/.
3. http://ir.shef.ac.uk/geoclef/2005/.
4. www.informatik.uni-trier.de/ ley/db/conf/gis/.

notamment les séries de conférences RIAO[5] [ARS 85], GIR et GIS qui voient associer la dimension temporelle à la dimension spatiale et/ou thématique : *spatio-temporal-textual search* respectivement, en 2004 [WID 04], en 2007 [MAR 07] et en 2010 [LIU 10].

De nombreux ouvrages de recherche abordent ces dimensions de la RIG. Nous pouvons citer les ouvrages *Georeferencing* [HIL 06], *The Geospatial Web* [SCH 07], *Linguistique et recherche d'information, la problématique du temps* [BAT 11], ou encore, les mémoires de thèses *Toponym resolution by text* [LEI 07], *Geographic aware web text mining* [MAR 08a], *Temporal information retrieval* [ALO 08], *Geographic information retrieval : classification, disambiguiation and modelling* [OVE 09], *Geographically constrained information retrieval* [AND 10], *Traitement automatique du langage pour l'indexation thématique et l'extraction d'informations temporelles* [KEV 11]. Ces travaux ciblent essentiellement la RIG dans des documents textuels ou multimédia de quelques lignes à quelques pages disponibles sur le *web*.

Les travaux présentés dans cet ouvrage ont, quant à eux, pour domaine d'application les bibliothèques numériques (BN) et plus particulièrement les corpus textuels. Nous pouvons citer le projet Google Books[6] avec plus de dix millions de livres numérisés à ce jour, le projet Bibliothèque numérique mondiale[7] avec, pour le moment, 6 142 objets numériques, le projet Europeana[8] avec dix millions d'objets numériques à ce jour, ou encore le projet Gallica[9] de la Bibliothèque nationale de France (BnF) avec plus d'un million de documents textuels (livres, périodiques, revues et journaux) numérisés. Comme beaucoup de bibliothèques et de médiathèques, la MIDR[10] de Pau Pyrénées numérise des documents de diverses natures (œuvres littéraires, récits

5. www.informatik.uni-trier.de/ ley/db/conf/riao/.
6. http://books.google.fr/books/.
7. www.wdl.org/fr/.
8. www.europeana.eu/portal/.
9. http://gallica.bnf.fr/.
10. Médiathèque intercommunale à dimension régionale de Pau Pyrénées : www.agglo-pau.fr/.

de voyage, journaux, cartes géographiques anciennes, lithographies, cartes postales, etc.) qui ont pour dénominateur commun de traiter d'un territoire restreint (les Pyrénées[11]), dans une période de l'Histoire déterminée (principalement du XVIII^{e} et du XIX^{e} siècles). Ce type de fonds documentaire contient d'abondantes références à l'histoire, à la géographie, au patrimoine, en somme au territoire [KER 11]. L'objectif visé dans le cadre de ces différents projets est d'offrir, au plus grand nombre, de nouveaux modes d'accès à des fonds documentaires désormais disponibles dans des formats numériques. Ainsi, ces projets mettent en œuvre des processus de marquage d'information, de construction d'index et d'interrogation *via* ces index.

Les documents qui composent le corpus de la MIDR sont particulièrement intéressants de par leur richesse en indications géographiques relatives au territoire pyrénéen. Des catégories d'utilisateurs de type « touriste », « élève », « pédagogue », « érudit » ou encore « bibliothécaire » ont été identifiées par les personnels de la MIDR. Ces utilisateurs visent l'exploitation du corpus documentaire *via* un système d'information adapté, en particulier, capable d'offrir des possibilités de recherche du point de vue du territoire décrit par ce corpus. Comme indiqué par Jihad Farhat et Luc Girard [FAR 04], systèmes de gestion électronique de documents (GED) et moteurs de recherche se complètent pour supporter les activités des usagers et professionnels des bibliothèques. Nous proposons d'étendre les fonctionnalités de ces systèmes par des services spécifiques dédiés au traitement des dimensions spatiales, temporelles ou thématiques de l'information. Ainsi, en comparaison avec les contenus du *web*, nous considérons que des fonds documentaires tels que ceux de la MIDR sont stables (le contenu d'un livre ne change pas au cours du temps) et homogènes pour autoriser des travaux d'indexation approfondis relatifs à chacune de ces trois dimensions.

11. http://fr.wikipedia.org/wiki/Pyrénées/.

I.2. De l'indexation d'information spatiale et temporelle à la recherche d'information multicritère

Les travaux de la littérature relatifs à la recherche d'information géographique dans des corpus textuels pointent notamment les verrous suivants :

– la reconnaissance et la résolution d'entités nommées spatiales et temporelles ;

– l'indexation spatiale et temporelle à des fins de RI ;

– l'appariement de couples document/requête et le calcul de scores de pertinence dédiés à la RI spatiale d'une part, et à la RI temporelle d'autre part ;

– la RI multicritère combinant les dimensions spatiales, temporelles et thématiques ;

– l'évaluation de tels systèmes de RIG.

Dans le laboratoire LIUPPA[12], au sein de l'équipe T2I[13], les travaux correspondants au premier point, menés sous la direction de Mauro Gaio [GAI 08], constituent le socle des travaux relatifs aux points suivants [PAL 12a].

Ainsi, la reconnaissance et la résolution d'entités nommées spatiales et temporelles [BAT 11, LEI 11] dans des documents textuels sont supportées par deux principales catégories d'approches. La première s'appuie sur un ensemble de règles, établies par des experts, permettant à un interpréteur de déterminer si un terme est une entité nommée ou non. La seconde, quant à elle, est basée sur un corpus d'apprentissage annoté manuellement permettant, à l'issu d'un traitement statistique, la constitution automatique de règles de découverte d'entités nommées

12. Laboratoire informatique de l'Université de Pau et des Pays de l'Adour : liuppa.univ-pau.fr/.
13. Traitement des informations spatiales, temporelles et thématiques pour l'adaptation de l'interaction au contexte et à l'utilisateur : http://liuppa.univ-pau.fr/live/EquipesdeRecherche/Equipe_T2I/.

applicables à des corpus plus larges. Conformément à la première catégorie d'approches, notre groupe de travail propose un ensemble de règles, bâties manuellement, dédiées à l'expression de l'espace et du temps dans un corpus composé de récits de voyages : ces règles permettent le marquage et une première interprétation symbolique des entités détectées (classification puis analyse d'éventuelles relations spatiales ou temporelles associées). Dans cette interprétation, nous avons distingué les entités absolues telles que « la ville de Pau » ou « l'an 2000 » des entités dites relatives telles que « les environs de la ville de Pau » ou « au début de l'an 2000 ». Rappelons que nous traitons uniquement les contenus textuels de documents, indépendamment de leur structure ou des métadescriptions associées.

L'indexation associe une interprétation numérique (géométrie, période calendaire) aux entités spatiales et temporelles détectées dans les textes. L'organisation des index peut, par exemple, dissocier complètement les références à l'espace et au thème dans des index indépendants ou bien combiner ces deux dimensions dans des structures spécifiques stockées dans un seul et même index [VAI 05]. Comme Paul Clough *et al.* [CLO 06], nous avons fait le choix de travailler sur des index spatiaux, temporels et thématiques indépendants (voir paragraphe 2.4.4). Les algorithmes d'interprétation des représentations symboliques des entités tiennent compte de l'aspect absolu ou relatif de leurs descriptions. La représentation numérique qui en résulte correspond au résultat d'une recherche dans des ressources de type gazetier dans le cas des entités absolues, par exemple.

L'appariement et le calcul de scores de pertinence ont également fait l'objet de nombreuses propositions en RI spatiale [AND 10] et en RI temporelle [ALO 08, BAT 11]. Comme pour la majorité de ces propositions, nous avons mis au point une RI spatiale et une RI temporelle supportées par des formules *ad hoc* adaptées à notre corpus.

La combinaison des dimensions spatiale, temporelle et/ou thématique en RIG est généralement mise en œuvre *via* des approches par filtrage [LIE 07, VAI 05]. Pour une plus grande expressivité dans le processus d'interrogation, nous avons proposé des opérateurs

d'exigence et de préférence qu'il est possible d'associer à chaque critère de recherche. Nous avons mis au point une méthode d'agrégation de résultats issus de différents systèmes de RI, tenant compte des niveaux d'exigence et de préférence exprimés dans la requête. Cette méthode est inspirée des approches d'agrégation mises en œuvre dans les systèmes d'aide à la décision [MAR 99] ainsi que dans les systèmes de recherche d'information multicritère [FAR 08].

La mise au point de premiers prototypes de RIG met en exergue la nécessité d'évaluer de tels systèmes [CAR 11, MAN 11]. Toutefois, à l'exception de campagnes telles que TEMPEVAL [VER 09] dédiée au temps ou GEOCLEF [GEY 05] dédiée au spatial et au thématique, il n'existe pas, à notre connaissance, de cadres d'évaluation de système de RIG combinant les dimensions spatiale, temporelle et thématique de l'information. Nous avons donc proposé un cadre expérimental dédié à ce type d'évaluation. Nous avons constitué une collection de tests et mis au point un protocole d'expérimentation que nous mettons en œuvre pour l'évaluation de nos prototypes.

L'ouvrage est organisé comme suit, afin de traiter ces différents axes de recherche. Nos propositions sont présentées dans quatre parties. Les deux premières parties visent l'indexation et la recherche d'informations spatiale et temporelle dans des corpus de documents textuels. Nous traitons de la RI spatiale, d'une part, et de la RI temporelle, d'autre part. Les deux parties suivantes, quant à elles, visent l'exploitation des index spatiaux et temporels ainsi produits dans un cadre de recherche d'information multicritère. Nous traitons ici pleinement de la RI géographique puisqu'il s'agit de RI combinant des critères de recherche spatiaux, temporels et thématiques.

L'indexation d'information spatiale et temporelle dans des documents textuels constitue le socle de ce travail. Dans cette partie indexation, la qualité de la reconnaissance et de la résolution d'entités nommées spatiales et temporelles est primordiale. Les trois parties suivantes exploitent les résultats de cette indexation à des fins de RI spatiale seule, de RI temporelle seule ou encore de RI multicritère combinant les trois dimensions géographiques.

Indexation d'information spatiale et temporelle dans des documents textuels

Dans cette partie, nous nous intéressons tout d'abord à la modélisation de l'information spatiale et temporelle dans un contexte de recherche d'information spécialisée et dédiée aux corpus textuels non structurés. Nous proposons un modèle spatial et un modèle temporel pivots [GAI 08, LES 06] intégrant les problématiques d'interprétation et de représentation des informations dans les index en vue de la mise en œuvre de calculs d'appariement en phase de recherche. Ensuite, nous proposons une première méthode d'extraction et d'indexation d'information spatiale, basée sur notre modèle pivot et sur un traitement sémantique spécifique [LES 06]. Nous adoptons une démarche similaire, pour proposer une méthode d'extraction et d'indexation d'information temporelle, basée sur notre modèle pivot et sur un traitement sémantique spécifique [LEP 07].

Recherche d'informations spatiale et temporelle dans des documents textuels

Dans cette partie, nous nous intéressons aux approches de recherche d'information mises en œuvre dans les systèmes dédiés à l'information spatiale et temporelle. Nous proposons une méthode de recherche d'information spatiale utilisant des fonctions de système d'information géographique (SIG) pour calculer des représentations géoréférencées et pour implémenter un calcul de pertinence spatiale [SAL 07a]. Selon une démarche similaire, nous proposons une méthode de recherche d'information temporelle dédiée [LEP 07].

Généralisation de représentations de données pour une recherche d'information multicritère

Nous avons fait le choix de traiter de manière spécifique chaque dimension de l'information géographique puis de les combiner dans des scénarios de recherche d'information. Afin d'éviter de possibles biais, il est important, avant toute combinaison, d'uniformiser les

représentations des données ainsi que les démarches de traitement des données relatives aux différentes dimensions. Nous proposons une démarche générique comparable à la généralisation par troncature ou lemmatisation de termes dans les approches de recherche d'information classiques. Ainsi, à partir des représentations indexées d'information spatiale et temporelle, nous construisons des index de plus haut niveau, appropriés à la mise en œuvre de modèles de recherche d'information éprouvés [PAL 10c, SAL 11].

Recherche d'information multicritère

Dans cette partie, nous nous intéressons aux approches de recherche d'information multicritère. Nous proposons de soumettre chaque critère d'une requête au système de recherche d'information de dimension spatiale, temporelle ou thématique approprié. Notons que, pour la dimension thématique, nous nous limitons aux approches mises en œuvre sur les termes en recherche d'information classique. Nous proposons plusieurs approches pour la combinaison des résultats issus de différents index et systèmes de recherche d'information. Nous proposons également, suivant le type d'utilisateur concerné, de nouveaux opérateurs dont le but est d'associer une plus grande expressivité à chaque critère de la requête et, par conséquent, d'améliorer la qualité des résultats [PAL 10b, PAL 10c, PAL 11, PAL 12a].

I.3. Organisation de l'ouvrage

Cet ouvrage s'articule de la manière suivante :

– le chapitre 1 présente le positionnement de ces travaux dans le domaine de la recherche d'information géographique ;

– le chapitre 2 est consacré à l'information spatiale et temporelle dans des documents textuels. Il décrit nos propositions relatives à l'indexation et à la recherche d'informations spatiale et temporelle dans des documents textuels ;

– le chapitre 3 traite de la généralisation de représentations de données avec pour objectif la préparation de la combinaison de résultats

issus d'une recherche d'information multidimensionnelle (spatiale, temporelle et thématique) et multicritère. Ce chapitre décrit nos propositions pour la recherche d'information multidimensionnelle et multicritère ;

– la conclusion est consacrée à un premier bilan puis présente un ensemble de perspectives comme autant de prolongements de ces travaux dans le domaine de la RIG.