Indexation sémantique et recherche d'information interactive

Le moteur GéoSem

Frédérik Bilhaut * — Franck Dumoncel * — Patrice Enjalbert * — Nicolas Hernandez **

** GREYC*
Université de Caen, Campus Côte de Nacre, boulevard du Maréchal Juin
BP 5186 14032 Caen Cedex

{fbilhaut, franck.dumoncel, patrice.enjalbert}@info.unicaen.fr

*** LINA-TALN CNRS FRE 2729*
2, Rue de la Houssiniere F-44322 Nantes Cedex 3 – FRANCE
nicolas.hernandez@univ-nantes.fr

RÉSUMÉ. *Parmi les différentes facettes de la recherche d'information en données textuelles, la recherche d'informations localisées dans l'espace et dans le temps constitue un domaine d'étude à part entière. Celle-ci nécessite en effet, pour l'indexation comme pour la recherche, des analyses linguistiques et des ressources spécifiques. Le projet GéoSem fut le cadre de conception de techniques d'indexation sémantique d'informations géographiques. Ces techniques se trouvent aujourd'hui mises en oeuvre au sein d'un moteur de recherche permettant la localisation intra-documentaire des informations, indexées selon des « axes sémantiques » géographiques (temps, espace et phénomène), sa généricité permettant de le paramétrer pour d'autres axes. Une visualisation de la localisation spatiale et temporelle de l'information est également proposée. Cet article décrit les différentes facettes du moteur.*

ABSTRACT. *Among the various facets of Information Retrieval in textual data, the search for information located in space and time constitutes a full research field. Indeed, it requires, for indexing as for retrieval, specific linguistic analyses and resources. The present paper roots in the GéoSem project, whose aim is to develop advanced, semantic-based methods for geographical documents retrieval. Today, these techniques are implemented within a search engine allowing an intra-document localisation of information, indexed according to "geographical" semantic axes (time, space and phenomenon). Moreover its generic and modular characteristics enable to cope with other « axes ». A visualisation of the spatio-temporal position of information is also proposed.*

MOTS-CLÉS : *recherche intra-documentaire d'informations géographiques, indexation sémantique (temporelle et spatiale), interface sensible aux besoins de l'utilisateur.*

KEYWORDS: *(text segments) information retrieval, geographical textual data, (time and space) semantic indexing, user-sensitive interface.*

CORIA 2007, pp. 65–76
Saint-Étienne, 28–30 mars 2007

1. Introduction

1.1. *Recherche d'information géographique*

L'information documentaire possède très fréquemment une « dimension géographique » en ce sens que les entités mentionnées (fleuves, villes, édifices publics…), les faits relatés (événements politiques, sportifs…) ou les observations décrites (de nature socio-économiques, par exemple) sont, d'une manière ou d'une autre, liés à une localisation dans un *espace géographique*. La notion de *recherche d'information géographique* (Larson, 2006), *spatiale* (Jones et al. 2002), ou *géographiquement informée* (*geographically-aware*) (Vaid et al., 2005) en découle comme domaine de recherche spécifique.

On s'accorde également à considérer qu'un système de Recherche d'Information (RI) ne doit pas traiter la localisation géographique sur le même plan que les autres « données » textuelles. Il y a à cela plusieurs bonnes raisons. Les unes sont de nature *terminologique et ontologique* : les noms d'Entités Géographiques (ou *géoréférencées, EG* dorénavant : villes, régions, fleuves…) présentent à la fois des propriétés de synonymie (plusieurs noms pour une même entité), d'ambiguïté (plusieurs entités pour un même nom), et d'imprécision (la localisation ne peut être déterminée en termes de coordonnées géographiques exactes). Les autres sont relatives à l'expression de la localisation *en langue*, qui ne se réduit pas à une nominalisation des *EG*, mais fait intervenir des relations spatiales (*près de, au nord/sud… de, entre… et …, dans,* etc.). Ces caractéristiques nécessitent des analyses linguistiques et des ressources (connaissances géographiques, ontologies) spécifiques pour l'indexation comme la recherche d'information subséquente. L'expérience du système GIPSY (Woodruff et al., 1994) est à ce titre significative.

De plus, des méthodes spécifiques, visuelles, d'interrogation peuvent être développées, dans lequel l'utilisateur « pointe » sur une carte les lieux et régions qui l'intéressent. Une navigation par l'espace (*spatial browsing*) dans la base documentaire peut être proposée (Larson, 2006).

Pour une part, ces remarques peuvent être reproduites en ce qui concerne le *temps*. Il est en effet fréquent que des faits, événements, ou observations exprimés dans un texte soient relatifs à une certaine date ou période temporelle. Les recherches sur l'indexation temporelle des informations textuelles sont particulièrement développées dans un contexte un peu différent de la RI « classique », à savoir l'Extraction d'Information (EI) (Pazienza, 1997) et pour des textes informatifs de type « dépêches d'agences » ou journalistique (Setzer, 2001). Un de ses développements notables est la définition d'un système d'annotation temporelle, formaté en XML, TIMEML[1]. Toutefois, la notion d'indexation temporelle peut tout à fait être développée en regard de tâches de recherche documentaire classiques (moins ambitieuses, donc, que l'EI, mais sur des corpus

[1] http://www.timeml.org/site/index.html

plus vastes). Et, comme dans le cas spatial, l'expression de la localisation temporelle répond à certaines règles spécifiques, nécessitant à la fois des connaissances, des analyses linguistiques, et des modes d'interrogations particuliers.

1.2. *Le moteur GéoSem*

Les expériences et réalisations présentées ici — menées dans le cadre du projet GéoSem (Enjalbert et Gaio 2006)[2] — s'inscrivent dans ce double contexte de recherche d'informations à la fois spatialement et temporellement qualifiées. Le contexte applicatif de GeoSem est celui de « l'Intelligence Territoriale »[3]. Il s'agit d'offrir à l'utilisateur des outils lui permettant d'analyser des situations locales, d'effectuer les diagnostics appropriés et finalement de prendre de bonnes décisions concernant les politiques d'aménagement du territoire. Cette spécificité implique un certain type de corpus, composé de textes expositifs, présentant des analyses socio-économiques dans lesquelles la variabilité spatiale et temporelle des phénomènes observés est un élément essentiel, comme dans l'extrait ci-dessous (Figure 1). Corrélativement, une requête naturelle portera sur un triple critère *Espace - Temps-Phénomène* : « Quelles informations puis-je avoir sur tel phénomène socio-économique, dans tel espace et dans telle période ? » (l'une ou l'autre des composantes étant évidemment susceptible de faire défaut). D'autre part, les documents traités sont fréquemment longs ou très longs (de quelques pages à quelques dizaines de pages pour certains rapports), ce qui fait que nous voulons retourner à l'utilisateur des *passages* des documents, de manière à permettre une *navigation* intra-documentaire.

> **De 1965 à 1985**, le nombre de collégiens et de lycéens a augmenté de 70%, mais selon des rythmes et avec des intensités différents selon les académies et les départements. Faible dans le Sud-Ouest et le Massif central, modérée en Bretagne et à Paris, l'augmentation a été considérable dans le Centre-Ouest, en Alsace, dans la région Rhône-Alpes et dans les départements de la grande banlieue parisienne où les effectifs ont souvent plus que doublé.

Figure 1. *Extrait de (Hérin et al., 1994)*

Le moteur de recherche GeoSem suit les principes usuels en matière de RI, distinguant une phase d'indexation « off-line » des ressources documentaires, et une phase d'interrogation « on-line » par un utilisateur, grâce à une interface appropriée. Mais il intègre aussi un certain nombre de spécificités, en réponse aux diverses contraintes relevées ci-dessus.

[2] Soutenu par le CNRS dans le cadre du programme interdisciplinaire « Société de l'Information » et par le Conseil Régional de Basse-Normandie dans le cadre d'une allocation postdoctorale.

[3] Des développements inspirés par le même projet initial, concernant la valorisation d'un patrimoine régional, sont menés à Pau au LIUPPA (Marquesuzaà et al., 2005)

(i) *Indexation multi-dimensionnelle*. Nous avons vu que les dimensions spatiales et temporelles ne peuvent être mises sur le même plan que les informations textuelles « ordinaires ». Les documents reçoivent donc une indexation selon trois *dimensions* ou *axes sémantiques* : Espace – Temps - Phénomène.

(ii) *Indexation sémantique*. La localisation spatiale et temporelle ne peut en aucune manière être traitée par une analyse « de surface » et une indexation par mots-clés. Une analyse sémantique, utilisant un ensemble de ressources linguistiques et de connaissances est nécessaire. Cette problématique est généralisée dans le moteur GéoSem : l'indexation est *a priori* fondée sur des analyses sémantiques et codées par des données symboliques, sous forme de *structures de traits*. Pour exploiter ces métadonnées, les comparer à la requête d'un l'utilisateur, des *comparateurs sémantique*s spécifiques à chaque axe sont définis.

(iii) *Indexation de passages*. Nous n'indexons pas les documents mais des passages de texte, identifiés par des *analyseurs discursifs*. Les métadonnées sont insérées dans le texte et liées à des passages pertinents. La réponse à une requête est une sélection de passages dans les documents sélectionnés.

(iv) *Formulation d'un besoin d'information en langage naturel*. La requête de l'utilisateur fait l'objet d'une analyse linguistique, par le même analyseur qui traite les ressources documentaires. Ce qui permet une expressivité plus riche que l'interrogation par mots-clés, et plus naturelle pour l'utilisateur non averti.

(v) *Navigation spatiale et temporelle*. L'interface utilisateur inclut un dispositif de visualisation sur une carte des localisations portées par le texte et la requête. Dans une version ultérieure, l'utilisateur pourra modifier à la souris la zone de requête et relancer le moteur avec ces nouvelles contraintes. Un dispositif similaire est proposé pour le temps avec une « frise temporelle » portant les périodes de la requête et des passages sélectionnés. Ce mode d'interaction est une réponse aux problèmes d'ambiguïté et d'imprécision sus-mentionnés.

(vi) *Un moteur modulaire, générique et interopérable*. Le moteur a été conçu de manière générique et indépendante des analyseurs et des comparateurs sémantiques. Il peut ainsi être adapté facilement pour prendre en compte les spécificités de nouveaux domaines ou corpus. La modularité est obtenue en considérant de manière séparée et distincte les phases de traitement : l'analyse linguistique et le calcul des index, le stockage des index et des ressources dans une base de données, les différents comparateurs sémantiques, le contrôleur global du moteur et l'interface utilisateur. L'interopérabilité est assurée en s'appuyant sur des échanges XML d'information, lesquels peuvent être exprimés en RDF.

Nous commencerons par décrire les principes de l'analyse des textes et de l'indexation. Nous présenterons ensuite l'interface interactive et terminerons en décrivant l'architecture du moteur, avant d'esquisser un bilan et quelques perspectives.

2. Analyse linguistique et indexation

Nous examinerons ici trois questions : en premier lieu les analyses spécifiques aux dimensions spatiale et temporelle ; puis le traitement de la dimension « phénomène » et la détermination des segments intra-documentaires à indexer.

2.1. *Indexation spatiale et temporelle*

L'étude de corpus a mis en évidence la structure sémantique de ces expressions, en relation avec son expression syntaxique. En gros, à partir de toponymes dénotant des entités géoréférencées, la langue procède par application d'un certain nombre d'opérateurs : opérateurs spatiaux (ou « géométriques ») d'une part tels que "le nord/sud… de", "le triangle X Y Z", " de X à Y " etc. ; et opérateurs de sélection d'entités au sein d'une zone donnée, selon divers types de critères : sociologiques, administratifs, physiques…

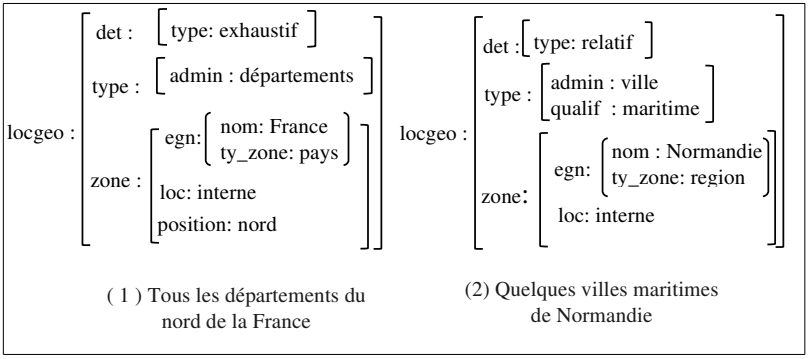

Figure 2. *Structures de traits produites par l'analyseur spatial*

L'indexation procède alors en deux temps, selon un schéma classique en compréhension automatique (Charnois et Enjalbert, 2005). Dans une première phase (*sémantique abstraite*) l'analyseur repère cette structure et la traduit en structures symboliques (structures de traits récursives, cf. figure 2). L'implémentation est réalisée en Prolog sous forme de grammaires DCG en utilisant les techniques GULP de Covington (Covington, 1994) pour les structures de traits, sur la plateforme LinguaStream[4]. L'analyseur prend en entrée le corpus préalablement étiqueté par un analyseur morphologique (actuellement le Tree Tagger (Schmid, 1994)). Par exemple pour l'espace, la grammaire comporte environ 160 règles syntagmatiques et un petit lexique créé manuellement d'environ 200 entrées, incluant des termes grammaticaux (déterminants, prépositions...) ainsi que

[4] http://www.linguastream.org

des termes administratifs (départements, villes, régions, etc.) et socio-économiques. Une base lexicale contenant plus de 100 000 lieux nommés français (*gazetteer*) est également utilisée comme ressource (Bilhaut et al., 2003a).

Cette représentation est très fidèle à l'expression en langue naturelle mais difficilement exploitable telle quelle par notre moteur de recherche qui doit être capable de comparer des représentations entre elles. Une seconde phase, dite *d'interprétation*, va la transformer vers une nouvelle structure dans laquelle toute expression spatiale se trouvera résumée sous la forme d'une entité géographique située dans un espace géographique géolocalisé. Une entité est caractérisée par trois champs : son type, son identifiant et son qualificatif. Son espace géolocalisé est une boîte rectangulaire représentée par la latitude et la longitude de ses coins nord-ouest et sud-est. Reprenons l'exemple précédent : « Les départements maritimes du nord de la France ». Nous obtenons :

spatial : (entité : (type : département, identifiant : libre , qualificatif : maritime)
boîte: (latitude nord-ouest, longitude nord-ouest, latitude sud-est, longitude sud-est))

Les différents champs de *l'entité* sont directement retranscrits de la structure « abstraite ». Les traits qui ne sont pas spécifiés sont qualifiés de « libres ». Les coordonnées figurant dans le champ boîte sont obtenues à partir de l'analyse du terme « nord de la France ». Pour ce faire, le système utilise un lexique d'entités nommées géographiques spécifiant leur géolocalisation en termes de « boîte englobante ». Un ensemble de règles permet de transformer une boîte pour représenter sa partie nord, sud, sa proximité, etc. Ces règles s'écrivent grâce à des opérateurs géométriques permettant, par exemple, d'étirer ou de déplacer une boîte ou de changer sa taille. Il est aisé de modifier ces règles pour expérimenter diverses façons d'interpréter des expressions[5].

L'analyseur temporel est conçu selon les mêmes principes. Les « points d'arrêt » sont ici les dates ; des opérateurs définissent des intervalles (« de X à Y », « entre X et Y », « les années X ») soumis à leur tour à une nouvelle classe d'opérateurs (« le début de X », « aux alentours de X »…). Un exemple d'expression traitée serait : « depuis le début des année 1950 jusqu'à la fin des années 1970 ». L'analyse linguistique fournit une représentation sémantique « abstraite ». L'interprétation référentielle produit une approximation de la période évoquée, représentée par un intervalle entre deux dates. C'est cet intervalle qui constitue l'indexation temporelle du texte.

[5] Ce calcul pose des problèmes de sémantique spatiale non triviaux (Charnois et al., 2003). La « boîte » proposée ici en constitue une approximation relativement grossière. L'interface visuelle interactive permet dans une certaine mesure de palier l'imprécision générée.

2.2. *Axe phénomène et indexation par passages*

Les analyses spatiale et temporelle établissent un premier jeu de métadonnées associées au document. Un troisième composant se charge alors de fournir un ensemble de « termes descripteurs » (au sens usuel en Recherche d'Information) caractéristiques de la dimension « Phénomène ». Rappelons également que nous souhaitons guider l'utilisateur vers des *passages* pertinents, relativement à sa requête, au sein des documents sélectionnés. Il nous faut donc aussi délimiter les segments textuels auxquels seront associées ces trois types d'informations.

Nous procédons *en premier lieu* à cette segmentation. Nous avons mis au point diverses méthodes dont il conviendra à terme d'évaluer la pertinence.

- La première consiste à exploiter le marquage typographique des documents, en prenant par exemple les paragraphes comme unité d'indexation. Il s'agit là d'une approche minimaliste, dans la mesure où l'homogénéité thématique de tels segments reste quelque peu approximative, et surtout parce que leur homogénéité sur les plans du temps et de l'espace est souvent inexistante.

- La seconde approche consiste à appliquer une méthode de segmentation thématique par cohésion lexicale (par exemple comme dans (Ferret et al., 2001)). On peut alors espérer un meilleur niveau d'homogénéité thématique, mais toujours aucune prise en compte spécifique des expressions spatio-temporelles.

- La troisième approche, plus spécifiquement développée dans le cadre du projet GéoSem, consiste à se fonder sur des modèles linguistiques d'ordre discursif, et en particulier sur l'hypothèse de l'encadrement du discours (Charolles, 1997). Ce modèle décrit des segments dits *cadres de discours*, homogènes par rapport à un critère sémantique (en l'occurrence une localisation spatiale ou temporelle) spécifié par une expression détachée en initiale de phrase dite *introducteur de cadre*. On voit immédiatement le bénéfice potentiel du repérage de telles structures dans le cas qui nous occupe ici, s'agissant de la cohésion temporelle ou spatiale des passages délimités. L'analyse automatique de ces structures constitue en soi un problème complexe, qui a fait l'objet d'une étude spécifique (Bilhaut et al., 2003b). Si le problème est encore loin d'être intégralement résolu, la qualité des résultats déjà obtenus autorisent amplement leur intégration au moteur ici présenté.

Dans un second temps, une fois établie la segmentation du document à indexer, nous procédons à l'analyse de la composante « phénomène ». La méthode actuellement retenue consiste à appliquer un calcul de type TF.IDF au niveau intra-documentaire afin d'isoler pour chaque segment les termes qui sont statistiquement saillants et discriminants. Ledit calcul est en soi parfaitement classique, à l'exception du fait que nous souhaitons obtenir ici des syntagmes plutôt que des mots, les premiers étant jugés beaucoup plus pertinents que les seconds pour indexer des documents liés à un domaine de spécialité.

3. L'interface

L'interface du moteur de recherche se présente sous la forme d'un formulaire à trois champs textuels : chacun désignant un axe sémantique d'interrogation (phénomène « quoi ? », spatial « où ? », temporel « quand ? »).

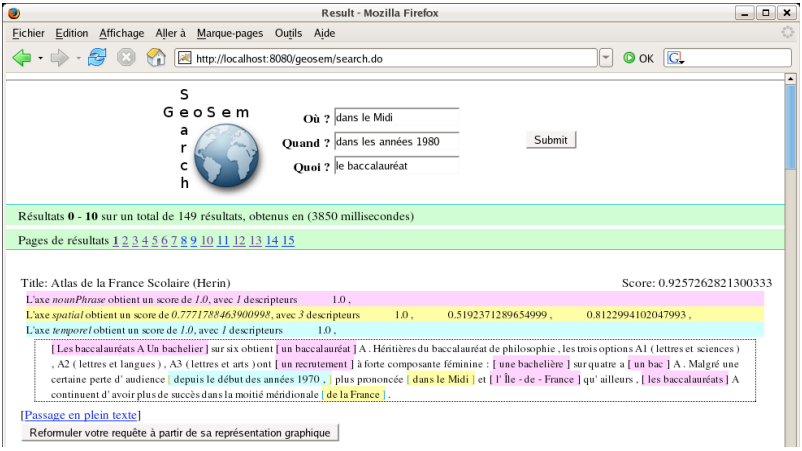

Figure 3. *Interface du moteur de recherche avec affichage d'un passage résultat*

Après formulation d'une requête et soumission au moteur, les passages résultats s'affichent au dessous du formulaire de la requête (la figure 3 illustre l'affichage du premier passage résultat pour la requête spécifiée en en-tête). Chacun de ces passages est présenté accompagné d'un certain nombre d'informations :

— Le titre du document dont il est issu et le score global qu'il a obtenu.

— Le détail de ce score pour chacun des axes suit cette première information.

— Le nombre de descripteurs pour chaque axe, ainsi que leur score individuel.

— La représentation sémantique des descripteurs telle qu'elle est produite par nos analyseurs (celle-ci est obtenue en survolant le score d'un descripteur).

— Le passage textuel, extrait du document.

— Un lien permettant une visualisation du passage dans son contexte.

_ Un lien permettant d'actionner le mode de reformulation graphique.

La reformulation graphique offre un moyen complémentaire de visualiser l'interprétation sémantique d'un passage et d'une requête. On propose à l'utilisateur une carte sur laquelle figurent, dans des couleurs différentes les boîtes qui « géopositionnent » les entités spatiales décrites dans la requête et trouvées dans le texte. L'axe temporel est représenté par une frise sur laquelle on trouve les différents

intervalles caractérisant les entités temporelles de la requête et des passages de textes.

L'utilisateur peut agrandir, rétrécir ou déplacer la boîte où le segment temporel caractérisant sa requête. Il est ainsi en mesure de reformuler ou de préciser sa requête. La figure 4 présente la requête « au nord du Massif Central » vis-à-vis d'un passage qui contient seulement un descripteur et qui désigne « au sud du Calvados ».

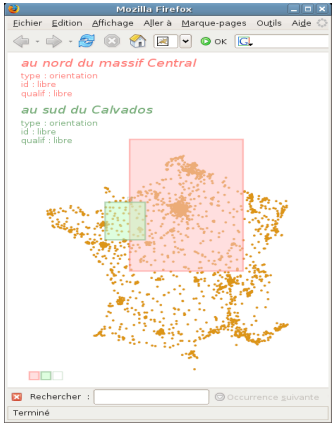

Figure 4. *Interface du moteur de recherche avec affichage d'un passage résultat*

4. Architecture

Comme annoncé plus haut, l'architecture du moteur GéoSem n'est pas spécifiquement liée à l'information géographique. Il s'agit au contraire d'une ossature générique dont la mise en oeuvre effective suppose un paramétrage décrivant les propriétés des « dimensions caractéristiques » de l'application considérée. Dans le cas de l'information géographique, il s'agit bien sûr des trois dimensions phénomène, temps et espace.

Compte tenu de la description de ces dimensions (nous détaillerons plus loin en quoi consiste précisément cette description), le système génère dynamiquement son interface utilisateur, et surtout adopte un procédé spécifique de recherche dans l'index. L'interface d'interrogation est composée d'autant de champs de saisie que le domaine comporte de dimensions. Le procédé de recherche est également spécifique à chaque domaine, le système autorisant le paramétrage complet de toutes les étapes.

4.1. *Structuration de l'index*

Ces différentes spécificités ont des implications fortes sur l'organisation de l'index lui-même. D'une part, le contenu informationnel de chaque unité documentaire doit être représenté par plusieurs entités distinctes, selon les différentes dimensions du domaine. D'autre part, la nature *sémantique* de l'indexation implique un second niveau de structuration, au niveau des descripteurs eux-mêmes, cette structure locale est elle-même dépendante de chaque dimension.

L'objectif d'*indexation de passages* a également un impact fort sur la nature de l'index. Les unités indexées sont en effet constituées de segments textuels qu'il convient de pouvoir délimiter au sein des documents. Le moteur est capable d'indexer n'importe quelle unité marquée selon le format XML de LinguaStream, sans hypothèse sur le schéma propre au document lui-même.

Compte tenu de ces propriétés, nous avons choisi de conserver l'ensemble des données au format XML. Leur insertion dans la base documentaire se résume alors à un filtrage des analyses linguistiques réalisées en amont, pour ne conserver que le marquage des segments effectivement indexés. Les index sont quant à eux représentés sous la forme d'un schéma XML spécifique, permettant d'associer à chaque identifiant de segment une série de structures de traits.

4.2. *Analyse des requêtes et processus de recherche*

Le processus global de recherche dans la base documentaire à partir d'une requête (en langue naturelle) inclut les étapes suivantes :

– Chaque composante de la requête est analysée en faisant appel à une chaîne de traitement spécifique. Chacune de ces chaînes de traitement renvoie une structure de traits représentant la valeur sémantique de la composante de la requête analysée.

– L'ensemble des passages présents dans l'index est consulté, et chacune de leurs dimensions est comparée à la structure correspondante de la requête.

– Finalement, les degrés de pertinence des différentes composantes de chaque passage sont combinés au sein d'une valeur unique qui permettra de les ordonner.

Sur ce point, le paramétrage du moteur passe par la définition d'une méthode de calcul de la pertinence pour chaque dimension. Pour l'axe temporel, le calcul de pertinence est binaire, par intersection ou non des deux intervalles. Pour le spatial, il fait intervenir à la fois un calcul de recouvrement des boîtes englobantes et une comparaison des structures administratives[6].

L'ensemble du système a été réalisé sous la forme d'une application Web J2EE (Java 2 Enterprise Edition). Le stockage des documents constituant l'index s'appuie

[6] http://www.info.unicaen.fr/~dumoncel/geosem/rapport_comparteur.pdf

sur la base de donnée XML native eXist, qui met à disposition des langages de requêtes XPath et XQuery pour parcourir les collections de documents.

5 Conclusion et perspectives

Le moteur de recherche GéoSem, conçu selon les principes que nous venons d'exposer est opérationnel et en phase d'expérimentation. La base documentaire actuellement traitée représente 66790 mots (425 Ko). Le temps de traitement d'une requête est principalement fonction du temps d'analyse sémantique des axes. Le temps de parcours de l'index (comprenant le temps des opérations de comparaison) varie de 1/20 à ¼ de seconde suivant l'axe[7]. Le temps d'analyse sémantique est environ de 1 seconde pour les axes phénomène et temporel et de 2 secondes pour l'axe spatial. Ces temps s'additionnent lors d'une requête multidimensionnelle. Il est important de rappeler que la plate-forme LinguaStream (qui réalise les analyses sémantiques) est avant tout un environnement pour le prototypage et l'expérimentation de chaînes de traitements en TAL et qu'aucune réflexion n'a été menée quant à l'optimisation de ses traitements. Par ailleurs dans l'état, chacune des analyses opérées est indépendante et par conséquent peut présenter des sous-traitements redondants que l'on pourrait factoriser. L'amélioration des performances est une première voie de nos recherches actuelles.

Une seconde question est celle du « passage à l'échelle » sur un plus vaste corpus. Le principal problème à résoudre ici concerne l'adaptation de l'analyseur spatial, qui demande l'acquisition de lexiques et d'ontologies spécifiques à certaines régions (maritimes, montagneuses…) ou à certains usages (critères sociologiques, démographiques… de sélection spatiale). Par ailleurs nous développons une interface interactive non seulement pour visualiser — sur une carte ou sur une frise — la localisation spatiale et temporelle de passages, mais aussi de spécifier ou de modifier une « zone de requête ».

Remerciements à toute l'équipe du projet GéoSem, dont le moteur de recherche ici présenté constitue en quelque sorte un aboutissement.

6. Bibliographie

Bilhaut F., Charnois T., Enjalbert P., Mathet Y., « Passage extraction in geographical documents », *Proc. Intelligent Information Systems 2003, New Trends in Intelligent Information Processing ans Web Mining*, Zakopane, Pologne, 2003a, p. 121-130,.

Bilhaut F., Ho-Dac, M., Borillo, A., Charnois T., Enjalbert P., Le Draoulec, A., Mathet, Y., Miguet, H., Péry-Woodley, M.-P., Sarda , L., « Indexation discursive pour la navigation

[7]Testé sur un Pentium 4 mono processeur cadencé à 2.80GHz avec 1GB de mémoire vive.

intradocumentaire : cadres temporels et spatiaux dans l'information géographique », *Actes de la 10e Conférence Traitement Automatique du Langage Naturel* (TALN), Batz-sur-Mer, 2003b, p. 315-320.

Charnois T., Mathet Y., Enjalbert P., Bilhaut F., « Geographic Reference Analysis for Geographic Document Querying », *Workshop on the Analysis of Geographic References, Human Language Technology Conference (NAACL-HLT)*, 2003, p. 55-62.

Charnois T., Enjalbert P., « Compréhension Automatique », in *Enjalbert P. (dir.), Sémantique et Traitement Automatique du Langage Naturel,* Hermes, 2005, Chapitre 7, p. 267-308.

Charolles M., « L'encadrement du dicours : Univers, champs, domaines et espace », Cahier de recherche linguistique, 6, 1997, p. 1-60.

Covington, M. A., *Natural language processing for prolog programmers*, Prentice hall, 1994.

Enjalbert P., Gaio, M « Traitements sémantiques pour l'information géographique, textes et cartes », *Revue internationale de Géomatique / European journal of GIS and Spatial Analysis*, Hermès Pub., 2006.

Ferret, O., Grau, B., Minel, J.-L., Porhiel, S., « Repérage de structures thématiques dans des textes », Actes de Traitement Automatique du Langage Naturel (TALN 01), Tours, France, 2001, p. 163-172.

Hérin, R., Rouault, R., Veshambre, V. *Atlas de la France scolaire. De la maternelle au lycée*, Collection Dynamiques du territoire, Reclus, 1994.

Jones, C.B., Purves,R., Ruas,A., Sanderson, M., Sester, M., van Kreveld, M., Weibel, R., « Spatial Information Retrieval and Geographical Ontologies. An Overview of the SPIRIT Project », *SIGIR'02*, August 11-15, 2002.

Larson, R.R., « Geographical Information Retrieval and Spatial Browsing », http://sherlock.berkeley.edu/geo_ir/PART1.html, 2006

Marquesuzaà, C., Etcheverry, P., Lesbegueries, J., « Exploiting Geospatial Markers to Explore and Resocialize Localised Documents », *GeoSpatial Semantics, First Int. Conf. GeoS 2005*, Springer LNCS 3799.

Pazienza M.T. (Dir.), Information Extraction, Springer Verlag, New York, 1997.

Setzer A., Temporal Information in Newswire Articles: an Annotation Scheme and Corpus Study, Thèse de doctorat, Université de Sheffield, Royaume-Uni, 2001.

Schmid, H., Probabilistic Part-of-Speech Tagging Using Decision Trees, Intl. Conference on New Methods in Language Processing. Manchester, UK , 1994.

Vaid, S., Jones, C.B., Joho, H., Sanderson, M., « Spatio-Textual Indexing for Gographical Search on the Web », Proc. 9th Int. Symp. on Spatial and Temporal Databases, Angra dos Reis, Brazil, 2005, p. 218-235.

Woodruff, A. G., Plaunt, C., « GIPSY: Geo-referenced Information Processing System ». *Journal of the American Society for Information Science, 45*, 1994, p. 645-655.