Modèle temporel, spatial et sémantique pour la découverte de relations entre événements

Arnaud Saval

UNIVERSITÉ de CAEN BASSE-NORMANDIE
UFR : Sciences
ÉCOLE DOCTORALE : SIMEM

THÈSE

Présentée par

Arnaud SAVAL

Et soutenue

Le 14 mars 2011

En vue de l'obtention du

DOCTORAT de l'UNIVERSITÉ de CAEN

spécialité: Informatique et applications
Arrêté du 7 août 2006

Modèle Temporel, Spatial et Sémantique pour la Découverte de Relations entre Événements

MEMBRES du JURY

Florence Le Ber	**Ingénieure en chef des Ponts, des Eaux et des Forêts, HDR**	**Université de Strasbourg**	**(rapporteur)**
Mauro Gaio	**Professeur**	**Université de Pau et des Pays de l'Adour**	**(rapporteur)**
Therry Saint-Gérand	**Professeur**	**Université de Caen**	
Abdel-Illah Mouaddib	**Professeur**	**Université de Caen**	
Maroua Bouzid	**Professeur**	**Université de Caen**	**directrice**
Stephan Brunessaux	**Chef du département IPCC**	**CASSIDIAN, EADS**	**co-encadrant**

Mis en page avec la classe thloria.

Table des matières

Remerciements

À ceux qui sont partis,
et surtout à ceux qui sont toujours là.

Je tiens à remercier ma directrice de thèse, Maroua Bouzid ainsi que mon encadrant Stephan Brunessaux sans qui cette thèse n'aurait pas vu le jour et je ne serais sans doute pas parvenu au bout de ce mémoire. Je tiens à les remercier plus particulièrement pour la compréhension et la patience qu'ils ont manifesté surtout pendant la dernière partie de cette thèse.

Je remercie les membres du jury pour s'être intéressé à mes travaux de thèse et pour avoir orienté mes travaux et m'avoir conseillé grâce à des discussions constructives tout au long de la thèse.

Je tiens à remercier les membres de l'université de Caen, dont les personnels du GREYC et du département informatique qui ont assuré que la thèse se déroule dans de bonnes conditions. Je tiens à remercier tout particulièrement Bruno Mermet, Gaële Simon et Bruno Zanuttini pour m'avoir transmis leur passion pour la recherche. Je tiens à remercier Abdel-Illah Mouaddib, les membres de l'équipe MAD et les doctorants du GREYC, passés et présents, pour m'avoir accueilli et supporté.

Je souhaite remercier les employés d'EADS qui ont fait tout leur possible pour faciliter mon intégration et plus particulièrement tous les membres de l'équipe IPCC dont la bonne ambiance et les discussions, parfois houleuses et interminables, se sont toujours révélées fructueuses. Je veux notamment remercier les personnes qui ont partagé le même bureau que moi pendant ces années de thèse ; Bruno Grilhères et Gérard Dupont pour leur disponibilité, leur bonne humeur et leurs conseils.

Je remercie aussi mes amis avec lesquels nous avons passé d'agréables moments de détentes, nécessaires pour prendre du recul. Ceux qui malgré l'éloignement m'ont toujours soutenu et répondu présent pour passer de bons moments ensemble. Je veux remercier ici Franck Facchetti pour son assistance et l'intérêt qu'il a montré à mes travaux de thèse malgré nos combats parfois violents.

Finalement, je remercie tous les membres de ma famille auprès de qui je n'ai pas pu être aussi disponible que je l'aurais souhaité. En particulier, je profite de cet espace pour dire merci à mon père, ma mère et mon frère Nicolas pour leur compréhension et leurs encouragements.

Introduction

Ce rapport de thèse décrit un travail de recherche dans l'étude d'un modèle temporel, spatial et sémantique dans le but de découvrir de nouvelles relations entre événements. Nous commencerons par définir quelles sont les informations que nous allons étudier ainsi que les différentes méthodes représentations possibles et adaptées afin de traiter automatiquement ces informations. Nous nous attarderons sur les principes de raisonnements liés à notre modèle. Nous terminerons par la présentation de résultats au travers d'un projet applicatif grâce auquel nous avons pu mener des expérimentations.

1 Contexte

Cette thèse a vu le jour grâce à la collaboration entre trois entités ; l'université de Caen, EADS et moi. Les problèmes à partir desquels les grandes lignes de cette thèse sont décrites se trouvaient être des besoins communs à chaque entité. L'étude des représentation des entités dans l'espace est un domaine de recherche qui s'inscrivait dans les travaux sur la modélisation engagés par l'équipe MAD de l'université de Caen. L'équipe IPCC chez EADS, cherchait à gagner en compétence en modélisation de l'information géographique à partir de sources ouvertes. Mon intérêt et ma motivation pour travailler sur des problèmes de modélisation formelle auxquels j'ai commencé à être confronté lors de mon stage de Master ont permis de débuter cette thèse.

L'équipe MAD s'intéresse depuis un certain temps à la modélisation du risque. Que ce soit par les projets étudiant avec la RobotCup Rescue, les thèses mises en place ou avec les différents ANR et projets régionaux qui traitent de la gestion du risque.

L'équipe IPCC se concentre sur les projets régionaux, nationaux ou européens de défense pour la DGA, l'EDA, les agences de renseignements ou plus généralement l'union européenne. Cette équipe est reconnue, entre autres, pour ses qualités en temps qu'intégrateur de systèmes. Cette particularité fait que les membres doivent constamment se tenir informer des nouveaux standards, normes et recherches publiés dans le domaine du traitement de l'information.

Plus particulièrement, un projet européen, nommé CITRINE dont EADS était partenaire, visait à surveiller les différentes catastrophes naturelles en passant par les alertes publiées sur Internet afin d'aider les Organisations Non Gouvernementales (ONG) à réagir plus efficacement en cas de crise en évitant d'être surchargé d'information. Ma participation à ce projet a grandement contribué aux travaux présentés dans ce rapport.

1.1 Diffusion de l'information sur Internet

La quantité et la qualité des informations contenues sur les blogs, forums et sites d'information s'accroissent de jour en jour. Ces sources concernent des domaines variés et se diffusent en temps quasi réel partout dans le monde, malgré un fort niveau de bruit. il existe de nombreuses causes de surcharge d'information sur Internet :

- L'augmentation rapide du taux de nouvelles informations en cours de production
- La facilité de reproduction et de transmission de données sur Internet
- L'augmentation dans les canaux entrants d'information disponibles (téléphone, e-mail, messagerie instantanée, rss)
- Les grandes quantités de données historiques sans intérêt
- Les contradictions et les inexactitudes dans les renseignements disponibles
- Le faible rapport signal-bruit
- L'absence d'une méthode de comparaison et de traitement de différents types d'informations
- Les informations échangées ne sont pas liées ou n'ont quasiment pas de structure globale pour caractériser les relations qui les lient

1.2 Web sémantique

Le Web sémantique est entièrement fondé sur le Web et ne remet pas en cause ce dernier. Le Web sémantique s'appuie donc sur la fonction primaire du Web « classique » : un moyen de publier et consulter des documents. Mais les documents traités par le Web sémantique contiennent non pas des textes en langage naturel (français, espagnol, chinois, etc.) mais des informations formalisées pour être traitées automatiquement. Ces documents sont générés, traités, échangés par des logiciels. Ces logiciels permettent souvent, sans connaissance informatique, de :

- générer des données sémantiques à partir de la saisie d'information par les utilisateurs ;
- agréger des données sémantiques afin d'être publiées ou traitées ;
- publier des données sémantiques avec une mise en forme personnalisée ou spécialisée ;
- échanger automatiquement des données en fonction de leurs relations sémantiques ;
- générer des données sémantiques automatiquement, sans saisie humaine, à partir de règles d'inférences.

1.3 Modélisation des événements

De nombreux travaux traitent de la modélisation de événements afin de représenter des phénomènes physique ou sociaux. Nous souhaitons représenter des phénomènes physiques grâce aux informations que tout le monde peut trouver sur Internet. Comment modéliser ces phénomènes, leurs évolutions ainsi que leurs impacts sur l'environnement ? De telles évolutions ont été étudiées dans plusieurs domaines tels que le comportement de foule [15], la migration d'animaux [84] ou encore la gestion de flotte de véhicules [166]. Nous voulons récupérer des nouvelles issues de pages web [176] et en extraire de l'information structurée (date, zone affectée, type de phénomène) grâce au Traitement Automatique des Langues Naturelles [87, 139]. De nombreux outils sont à notre disposition, cependant comme nous le verrons, leurs résultats peuvent varier d'un domaine à un autre.

2 Contributions

2.1 Motivations

Nous nous intéressons à la découverte de relations parmi l'information non structurée grâce à une modélisation temporelle, spatiale et sémantique. Pour inférer automatiquement ces relations, nous avons besoin d'une base de connaissance structurée suivant une modélisation formelle. L'intérêt est d'enrichir les modélisations temporelles et spatiales avec les apports d'une description

sémantique des informations afin d'obtenir de nouvelles règles de découverte de relations.

2.2 Contributions

Notre approche vise à tirer partie de la structure sémantique d'informations disponibles sur Internet afin de les combiner avec des propriétés temporelles et spatiales. Une fois structurée selon une ontologie et à l'intérieur d'une base de connaissance, l'application de méthodes de raisonnement sur ces informations vont mettre en évidence des liens de causalités et d'effets pour représenter une situation donnée. Nous pouvons séparer les contributions dans cette thèse en quatre grandes parties :

- Proposition d'un modèle de représentation des événements et leur relations à partir de critères temporels, spatiaux et sémantiques
- Découvertes de nouvelles relations
- Proposition de méthodes de découverte de contextes d'événements
- Application aux catastrophes naturelles (CITRINE)

Nous reprenons ces contributions point par point plus en détail.

Nous avons étudié les différentes possibilités de représentations temporelle, spatiale et sémantique des événements afin de proposer une modélisation formelle de ces informations à traiter. Ces représentations nous ont permis de constituer un modèle de représentation des événements et de leurs relations. L'étude de ce modèle passe par l'examen des impacts de la représentation sémantique sur les aspects de la modélisation temporelle et spatiale des événements. Nous nous efforçons d'imposer le moins d'hypothèses possibles pour fournir un modèle qui soit assez générique pour être appliqué à différents domaines (dont la gestion des risque).

Ce modèle constitue une base importante des travaux présenter dans cette thèse. En effet, la représentation formelle des éléments du modèle permet de mettre en place des règles de découvertes de nouvelles relations. Ceci est dû aux propriétés intéressantes qui découlent des contraintes supportées par les représentations temporelle, spatiale et sémantique des éléments du modèle.

Nous présentons également des méthodes de découvertes de contexte afin de construire semi-automatiquement les connaissances qui constitueront le contexte des événements d'un domaine considéré. Nous détaillons quelles sont les informations nécessaires et les méthodes pour construire ces éléments de connaissances dépendant du domaine.

Finalement, grâce à un projet européen de gestion de catastrophes naturelles (CITRINE), nous avons pu appliquer nos précédentes contributions afin de les confronter à des données utilisateurs issues de flux d'informations. Depuis juin 2008, date de la dernière version majeure et stable du modèle, nous avons mis à disposition une application, Agate, qui a été testée et accueillie positivement par des utilisateurs finaux et reste disponible pour à la fois présenter les précédentes contributions mais également constituer un corpus de données pour de futurs travaux.

3 Annonce du plan

Le mémoire est composé de deux parties :

- les deux premiers chapitres constituent la première partie qui rappelle l'état de l'art
- la seconde partie, composée des trois chapitres suivants, est consacrée aux contributions

Le premier chapitre introduit la notion de représentation de l'information en décrivant plus particulièrement les théories majeures de l'information mais également en présentant les ca-

ractéristiques de l'information dans les domaines de l'extraction d'information et la recherche d'information. Il se conclut par le rappel des descriptions sémantiques de l'information.

Le deuxième chapitre présente les travaux existants sur les notions de représentation du temps et de représentation de l'espace. Nous nous attarderons également sur la représentation des relations temporelles et des relations spatiales entre entités.

Le troisième chapitre présente les éléments qui servent de socle au modèle de représentation des événements que nous proposons. Tout d'abord, nous introduisons la notion d'événement puis nous définissons leurs évolutions en étudiant la notion de phénomène.

Le quatrième chapitre étudie les possibilités de raisonnement offert par le modèle. Cette étude débute par la présentation des méthodes de découverte automatique de contexte. Ensuite, nous nous intéressons aux règles d'agrégation d'événements qui déterminent la découverte de relations non évidentes entre les événements et les phénomènes.

Le cinquième chapitre traite de l'application du modèle à un cas concret de surveillance des catastrophes naturelles. Nous décrirons la spécification et les impacts de l'architecture que nous avons employée pour expérimenter le modèle de représentations des événements. Nous nous attarderons également sur les procédures d'expérimentation et l'analyse des résultats obtenus qui donnera lieu à une discussion sur le modèle.

Dans le dernier chapitre nous concluons et dressons un bilan de nos contributions et présentons les perspectives de ces travaux.

Première partie

État de l'art : Représentation de l'information

Chapitre 1

Représentation de l'information

Sommaire

Ce chapitre présente les notions de base introduisant la représentation de l'information à laquelle il sera fait référence régulièrement au long de ce mémoire.

Nous aborderons les notions d'information, de sources d'information ainsi que la notion de traitement de l'information en vue de sa structuration, enfin nous introduisons succinctement la notion de représentation de l'information sémantique.

1.1 Information

Cette partie s'attarde sur les théories majeures qui cherchent à définir ce qu'est l'information ; la théorie de l'information de Shannon et la théorie algorithmique de l'information. Le but n'est pas de faire une étude précise de chaque théorie pour les mettre en concurrence mais de présenter la notion d'information au travers de deux points de vue complémentaires.

1.1.1 Théorie de l'information de Shannon

L'origine du concept d'information est une notion apparue lors des recherches théoriques sur les systèmes de télécommunication. Ces recherches, menées par Boltzmann et Markov, étudiaient la notion de probabilité d'un événement et sa mesure. Ces travaux posèrent les bases de la théorie de l'information de Shannon qui l'explicita dans son article *A Mathematical Theory of Communications* [201] publié en 1948.

i) Définition

Pour Shannon, une information est uniquement définie par sa probabilité. Ce qui revient à confondre l'information et la mesure de l'incertitude à partir de la probabilité de l'événement. Tout lien avec le sens de l'information ou le moyen de transport de l'information est mise à l'écart pour se focaliser sur une représentation mathématique.

ii) Mesure

Une unité d'information est appelée bit. Un bit peut se voir affecter seulement deux valeurs 0 ou 1. L'affectation d'une valeur à un bit va fixer la probabilité de l'événement et ainsi va directement impacter la mesure de l'incertitude.

iii) Exemple

le code binaire suivant 001100 010010 011110 100001 101101 110011 mesure 36 bits

iv) Conclusion

La théorie de l'information de Shannon introduit une représentation probabiliste de l'information qui a donné lieu à de nombreuses applications dans des domaines divers :

- informatique [228],
- télécommunications [202],
- traitement de signal [171],
- sciences humaines [149]

Si cette représentation probabiliste de l'information a mené à de grandes avancées qui restent encore en pratique aujourd'hui, les mesures seules du contenu de l'information bien que nécessaires ne sont pas suffisantes pour rendre compte de toute la complexité inhérente à l'information transmise.

1.1.2 Théorie algorithmique de l'information

L'intérêt de cette théorie est de définir formellement les relations entre la complexité de l'information et son contenu. Gregory Chaitin a résumé la théorie algorithmique de l'information comme "*the result of putting Shannon's information theory and Turing's computability theory into a cocktail shaker and shaking vigorously*". Les premières bases ont été posées par Solomonov [205] dans les années 60 pour être développées par Kolmogorov et Chaitin quelques années plus tard.

i) Définition

Cette théorie étudie les liens entre la **description** et la **complexité** d'un objet. À la différence de la théorie de Shannon, ce n'est pas seulement le contenu de l'information qui est étudié mais également la façon dont elle est décrite.

ii) Mesure

La mesure est à la fois qualitative et quantitative. Un algorithme va définir la **description** d'un objet et la **complexité** est définie par la taille de l'algorithme et son temps de calcul. Le contenu de l'information correspond alors à la complexité de l'objet. Pour donner une idée intuitive de la mesure de la complexité, plus un objet est complexe à écrire, plus il contient d'information (plus un mot est long, plus il contient de lettres).

iii) Exemple

Supposons que nous voulions décrire à quelqu'un la chaîne de caractères suivante :

00001111

Il existe plusieurs façons de donner sa description :

- un 0, suivit d'un autre 0, suivit d'un autre 0, suivit d'un autre 0, suivit d'un 1, suivit d'un autre 1, suivit d'un autre 1, suivit d'un autre 1
- quatre 0 suivit de quatre 1

Même si il s'agit du même objet, les longueurs des descriptions sont différentes. Cette mesure n'est pas forcément la plus adaptée pour définir l'information transmise. La seconde description évite de répertorier l'ensemble des éléments qui caractérisent l'objet en se servant des propriétés de répétition des éléments dans l'objet.

iv) Conclusion

La théorie algorithmique de l'information crée un lien entre la théorie de l'information de Shannon et l'application à la notion de machine de Turing. Cette théorie s'est surtout vu utilisée dans les domaines suivants :

- la biologie [38],
- la physique [33],
- la philosophie [80]

La théorie de l'information de Shannon et la théorie algorithmique de l'information que nous venons de présenter formalise la notion d'information telle qu'elle sera utilisée tout au long de ce manuscrit. Cependant, la notion de représentations de l'information que nous introduisons ne se limite pas à décrire les objets et leurs complexités mais nécessite un traitement plus large de la notion d'information.

1.2 Source d'information

Représenter l'information passe évidemment par une description de son contenu mais une information seule est souvent limitée sans son contexte [182]. Une partie de ce contexte peut être directement liée à la source d'information. Il existe de nombreux types de sources d'information qui varient aussi bien en terme de présentation que de moyens pour y accéder. En fonction de la nature de ces sources, les informations sont interprétables différemment [19]. C'est pourquoi nous allons nous intéresser aux caractéristiques particulières inhérentes aux sources d'information. L'étude de la représentation de l'information est dépendante de l'étude de la source d'information qui l'a publiée.

1.2.1 Contexte

La phrase « L'alcool bout à 176° » paraît se suffire à elle même. Cependant, en apportant plusieurs éléments de contexte concernant sa source, l'interprétation de l'information contenue dans la phrase précédente est totalement différente :

Cas 1 : cette phrase a été traduite à partir d'un dictionnaire de chimie écrit en allemand

Cas 2 : cette phrase a été écrite lors d'un cours de sciences physiques où l'alcool est précédemment défini comme un alcool avec chaîne saturée non ramifiée et 1-ol contenant 7 carbones

Cas 3 : cette phrase a été trouvée sur un site internet après une recherche des termes « alcool » et « distillation »

Le contexte est essentiel car il va qualifier l'information vis à vis de son destinataire et de sa source. Dans le premier cas, nous supposons que les 176 degrés sont des degrés Fahrenheit alors que dans le second cas, il s'agit sans doute de degrés Celcius. Enfin dans le dernier cas, le contexte n'est pas forcément clair même si la langue du site pourrait nous aiguiller. Le type de la source d'information va alors modifier la façon dont celle ci est perçue.

1.2.2 Typologie

La mise en place d'une typologie des sources d'information est primordial afin de les catégoriser. En effet, savoir qu'une source d'information a des similitudes (de nature, d'accès ou de thème d'intérêt) avec une seconde source d'information nous donne des indices qui vont faciliter son interprétation grâce aux indications déjà détectées dans la seconde source d'information.

1.2.2.1 Catégories

Qu'il s'agisse d'un cycle de veille, de construction d'un thésaurus ou dans le domaine de la recherche d'information, les sources d'information subissent d'abord un classement sur trois critères[115] :

1. sources **formelles** (documents écrits), **informelles** (entretiens, observations)
2. sources **internes** (notes, mémo internes), sources **externes** (publications, compte, dossier de presse)
3. sources **ouvertes** (accessible facilement), sources **fermées** (payantes, demande des investigations),

En reprenant notre exemple précédent, nous pouvons catégoriser la source d'information grâce au contexte :

Cas 1 : nous sommes en présence d'une source d'information **formelle**, **externe** et **fermée.** En effet, ce dictionnaire est un document écrit et publié qui cible les spécialistes et les bibliothèques qui devront payer pour l'obtenir.

Cas 2 : la source d'information est **informelle**, **interne** et **ouverte.** Il s'agit de notes dictées par un professeur lors d'un cours de chimie, elle est donc accessible à tous les étudiants.

Cas 3 : la source d'information est **formelle**, **externe** et **ouverte.** La page web contenant la phrase est un document structuré publié sur Internet et accessible sans restriction.

De plus, l'information est elle aussi divisée en trois catégories qui indiquent son moyen d'obtention :

- information **blanche** *:* librement accessible,
- information **grise** : accès payant, résultats d'investigation,
- information **noire** : obtenue illégalement, espionnage

1.2.2.2 Nature des sources

Selon les besoins de la typologie, il est également possible de séparer les sources d'information en fonction de la nature de leurs supports :

Textuel : journaux, livres, courrier

Sonore : radio, musique, podcast

Visuel : photographies, prospectus

Multimédia : à la fois textuel, sonore et/ou visuel tels que télévision, jeux vidéos, discussion entre personnes

Hypertexte : extension du support multimédia afin de lier entre elles les informations ; pages web.

Il faut remarquer que le fait d'appartenir à l'une ou l'autre des catégories n'entraîne pas de restrictions sur la nature possible des supports des sources d'information.

1.2.2.3 Intérêt des sources ouvertes

Dans ce manuscrit, nous allons nous intéresser plus particulièrement aux sources ouvertes. Les sources ouvertes désignent les média accessibles sans restriction de droits (Internet, bases de données publiques, journaux, cd-rom, télévision, radio). Elles fournissent d'importants volumes de données multimédia hétérogènes (images, texte, audio, vidéo, ...) nécessitant des traitements adaptés. En plus de nous intéresser aux problèmes posés par cette hétérogénéité, nous nous intéresserons également à la structuration de ces données pour les rendre interprétables par une machine. Comme nous allons le voir plus en détails, la plupart des méthodes impliquent des enchaînements de traitements qui intègrent des algorithmes capables de traiter automatiquement ces problèmes.

Cependant, les avantages des sources ouvertes impliquent également des inconvénients ; par exemple, Internet contient toutes sortes de données et elles sont présentées sous une multitude de formes : blog, forum, tweet, facebook. Nous sommes face à un type d'information volatile, d'une précision variable et disponible dans des structures diverses[19].

1.2.3 Caractéristiques

Nous disposons maintenant de moyens pour identifier la source d'information. Nous pouvons alors étudier les caractéristiques propres à chaque sources. En effet, en fonction de son type, la source d'information dispose de propriété variables concernant sa mise à jour, sa pertinence, sa disponibilité et sa fiabilité.

1.2.3.1 Mise à jour

La mise à jour désigne la notion de fraîcheur de l'information. Étudions la date de mise à disposition selon chaque cas sur notre exemple :

Cas 1 : le dictionnaire de chimie a été écrit en 1811

Cas 2 : les notes de cours ont été prises durant l'année 2009

Cas 3 : la page web a été publiée en 2010

La date de mise à disposition n'est qu'un des éléments permettant de définir la fraîcheur de l'information ; la date de création, la date de modification, la date d'archivage, la date d'accès, ... La mise à jour de la source d'information donne des renseignements pour enrichir le contexte de l'information.

1.2.3.2 Pertinence

La pertinence correspond à la valeur de l'intérêt donné à la source d'information par une personne. En général, plus la pertinence d'une source d'information sera élevée plus les informations publiées par cette source se révéleront d'intérêt[79]. La pertinence d'une information varie en fonction du besoin de recherche d'information. Étudions le point de vue d'un historien et d'un étudiant en chimie sur chaque cas de notre exemple ;

Cas 1 : le dictionnaire peut être utile à l'historien, l'étudiant en chimie sera moins intéressé

Cas 2 : les notes de cours seront sans objet pour l'historien mais elle seront utiles à l'étudiant lors de ses révisions

Cas 3 : le site web n'a pas d'intérêt ni pour l'un, ni pour l'autre car il existe d'autres sites web mieux structurés et mieux adaptés à leurs besoins

La pertinence de la source d'information est indissociable du besoin d'une recherche d'information. Même si l'information comparée est la même (dans notre exemple « l'alcool bout à 176° »), la pertinence de la source d'information va solliciter l'intérêt de la personne et donc l'interprétation de l'information.

1.2.3.3 Disponibilité

La disponibilité désigne la notion d'accessibilité de la source d'information. Étudions la disponibilité des sources d'information pour chaque cas de notre exemple ;

Cas 1 : le livre est disponible depuis 1811 dans de nombreuses bibliothèques et fait maintenant partie du domaine publique

Cas 2 : les notes de cours seront accessibles à l'étudiant pendant son année scolaire, voire pendant plusieurs années ensuite selon ses capacités d'archivage

Cas 3 : le site web n'est pas référencé pour être très stable, il arrive parfois qu'il soit impossible de s'y connecter

En fonction de sa nature, une source d'information a un cycle de vie différent[192]. L'accès à une page web est plus aléatoire que l'accès à un livre dans une bibliothèque. La disponibilité de la source d'information est un des indicateurs de l'intérêt de l'information contenue. Ainsi plus une information est facilement accessible, plus elle pourra être facilement diffusée.

Nous devons remarquer que ce n'est pas parce qu'une information est difficilement accessible qu'elle n'est pas d'intérêt [32]. Au contraire, dans certains cas (rupture de stock d'un livre, surcharge de serveur web ...) la difficulté d'accès à la source est un indice supplémentaire de l'intérêt porté à cette information.

1.2.3.4 Fiabilité

La fiabilité désigne la notion de crédibilité accordée à une information et/ou à sa source. Reprenons notre exemple et attardons nous sur l'auteur pour chaque cas ;

Cas 1 : le dictionnaire a été écrit par Martin Heinrich Klaproth qui était professeur à l'université de Berlin et le "père de la chimie analytique"

Cas 2 : les notes de cours ont été écrites par un étudiant en chimie

Cas 3 : le site web est publié de manière anonyme

Dans le premier cas, il semble assez naturel d'accorder sa confiance à l'auteur et par conséquence à la source d'information. Nous sommes tentés d'interpréter l'information comme véridique puisqu'il s'agit d'un document réalisé par un expert dans son domaine. Dans le seconde cas, le cours a été donné par un professeur de chimie, il semble naturel d'accorder du crédit à ces notes. Il faut pourtant souligner que ce n'est pas le professeur qui a écrit ces notes mais c'est l'étudiant qui a retranscrit le cours. Les notes ne reflètent pas complètement le cours. Dans le troisième cas, nous n'avons aucune information concernant l'auteur qui a souhaité rester anonyme, il existe bien sûr des moyens d'accéder à l'identité de cette personne mais aucun qui ne soit immédiat et certain.

La connaissance de la fiabilité de l'auteur n'est qu'un élément permettant de définir le critère de fiabilité. En effet, le problème de fiabilité de l'information constitue une partie du problème plus général de cotation de l'information [1, 79] qui se rencontre lors de la veille (veille économique, veille informationnelle, renseignement).

Dans nos travaux, nous cherchons à représenter l'information, plus particulièrement l'information disponible et accessible à n'importe qui sur Internet. Dans la section suivante, nous allons nous attarder sur les avantages et les inconvénients des sources ouvertes liées à Internet.

1.2.4 Particularités des sources ouvertes sur Internet

La mise à jour de ces sources ouvertes est une des caractéristiques intéressantes de ce type de source. En effet, l'apparition et l'évolution d'une information est facilement et rapidement diffusée (en s'appuyant sur les réseaux sociaux), ce qui permet d'être au courant d'une information peu de temps après son apparition. Cette fréquence et cette fraîcheur de l'information ont été des éléments déterminants lors du choix du type de sources que nous avons été amenés à exploiter dans nos travaux.

La pertinence de ces sources ouvertes est plus discutable. En effet, l'intérêt de l'information publiée varie en fonction des sites mais également entre les pages web d'un même site. Toutes les pages de Wikipedia [63] n'ont pas le même intérêt. Les articles publiés par les sites de news (reuters.com, lemonde.fr [1]), par exemple, sont écrits par des journalistes avec leur spécialités respectives.

La disponibilité de ces sources ouvertes peut se voir restreinte. En effet, il peut arriver que des problèmes liés aux réseaux ou aux serveurs empêchent l'accès à certaines informations. De plus, certaines censures peuvent être appliquées au niveau des gouvernements pour limiter l'accès à certaines sources d'information

. Pourtant ces éléments ne sont pas un inconvénient car il existe de nombreux moyens de résoudre ces problèmes (mis en place de serveurs miroirs, utilisations de proxy/VPN/réseaux d'anonymisation). La facilité de publication et de diffusion d'une information sur Internet compense, le plus souvent, les inconvénients liés aux limitations techniques.

La fiabilité de ces sources ouvertes est liée au problème de la cotation de l'information disponible en sources ouvertes qui vise à définir quels sont les moyens et dans quelle mesure accorder sa confiance à une information. Nous ne rentrerons pas plus dans les détails de ce problème car cette étude dépasse le simple cadre de cette thèse. Cependant, nous supposerons dans le reste de ce manuscrit que l'information que nous rencontrerons est fiable sauf dans certains circonstances particulières où nous indiquerons explicitement que ce n'est pas le cas (par exemple lors du traitement de la désinformation ou de l'ambiguïté).

D'autre part, les sources ouvertes sur Internet constituent une source d'information en augmentation constante. Pour répondre à un besoin d'information précis, une recherche d'information sur Internet est devenue une tâche difficile tant cette somme d'information est gigantesque et versatile. Nous sommes en présence d'une surcharge d'information. L'information disponible dans ce type de sources ouvertes est une information brute et sans structure (sauf celle des règles de grammaires et de syntaxe de la langue du texte). La plupart des moteurs de recherches tendent à répondre à ce problème de surcharge en indexant par termes les informations publiées sur internet et en proposant un moyen centralisé pour rechercher de l'information qui présentera les résultats sous forme de liens vers les sites d'intérêts. Cependant, cette forme de représentation de l'information par les moteurs de recherche néglige, dans la plupart des cas, la structure et la sémantique de l'information proposées dans le but d'améliorer l'efficacité de la recherche d'information par mots clé. Cette représentation, bien qu'efficace, n'est pas appropriée pour la découverte de relations entre les informations. Pour obtenir ces informations de structures et de sémantiques qui nous intéressent, nous allons avoir besoin d'effectuer un traitement préalable sur ces informations brutes.

1.3 Traitement de l'information

Nous nous intéressons aux informations disponibles dans les sources ouvertes sur Internet. Ces sources d'information constituent de grands ensembles qui posent de nombreuses difficultés [117]. Du traitement de grands ensembles d'images à la détection de signaux faibles noyés dans les masses d'information, de plus en plus fréquemment, les sources ouvertes demandent d'adapter des techniques particulières pour effectuer un traitement spécifique. Le traitement de l'information sur Internet se déroule habituellement en cinq étapes :

1. Collecte des données
2. Normalisation des données

3. Extraction des entités d'intérêt
4. Analyse des entités
5. Indexation

Afin de pouvoir suivre l'apparition et l'évolution de nouvelles données à partir de ces sources ouvertes, nous devons répéter continuellement ces étapes. Ce cycle permet de convertir l'information brute en éléments de connaissance et de les indexer afin de pouvoir les réutiliser par la suite.

1.3.1 Collecte de données

La collecte des données sur Internet consiste principalement à parcourir les différents sites web en suivant les liens présents dans les pages. Il existe plusieurs stratégies de collecte[13, 77, 164, 173] :

- collecte exploratoire
- collecte régulière
- collecte personnalisée

1.3.1.1 Besoin de collecte

Il est nécessaire de définir son besoin avant de commencer toute collecte. En effet, le très grand nombre de sources d'information fait que la collecte complète de toutes les informations est difficilement réalisable. Ce besoin est souvent caractérisé par la volonté d'enrichir notre connaissance sur un domaine ou une entité. Nous avons alors à notre disposition un premier ensemble de connaissances (base de connaissances) que nous souhaitons enrichir : un exemple de besoin pourrait être de se tenir au courant des nouvelles dans le monde.

1.3.1.2 Collecte exploratoire

La collecte exploratoire est surtout utilisée pour découvrir de nouvelles sources d'information. Elle consiste à parcourir sans arrêt les liens entre les pages web en détectant les nouvelles pages web et à parcourir en priorité ces nouvelles pages.

1.3.1.3 Collecte régulière

La collecte régulière ne va pas chercher de nouvelles sources d'information mais va essayer de détecter les changements sur un ensemble de site web prédéfini. Lorsqu'une page est mise à jour, les nouvelles informations qu'elle contient seront récupérées par la collecte mais celle-ci ne suivra pas les nouveaux liens hypertextes apparus dans ces pages.

1.3.1.4 Collecte personnalisée

La collecte personnalisée est une sorte de compromis entre les deux derniers types de collecte. En effet, celle ci permet à la fois une collecte exploratoire et régulière. Ce type de collecte est le plus efficace pour répondre à un problème de veille d'information. nous pouvons alors facilement limiter la collecte à certaines langues mais également à certains sites (forums, blog, news).

Cette collecte personnalisée permet à la fois de garder l'hétérogénéité présente dans les informations trouvées sur le web tout en définissant des critères sur les informations brutes qui

seront récoltées. Ceci est évidement nécessaire pour appliquer efficacement les prochaines étapes du cycle de traitement de l'information.

1.3.2 Normalisation

L'étape de normalisation consiste à transformer l'information brute collectée dans un format adapté en vue des futurs traitements. La normalisation est dépendante à la fois de la nature de la source d'information mais également du format de destination de cette information [179]. En général, ce format de destination est un standard qui permet d'appliquer les processus de traitement indépendemment des contraintes et limites de l'information brute. Plus particulièrement, nous allons nous décrire succinctement les principaux formats d'échange que sont le HTML, XML et GML. Ces formats d'échange de données nous seront utiles dans la suite de ce manuscrit pour modéliser l'information disponible en sources ouvertes.

1.3.2.1 HTML

Le principal format d'échange de document sur Internet est le *HTML* pour *HyperText Markup Language*. Ce langage de balises a été proposé par Tim Berners-Lee en 1991[17] pour transmettre un ensemble de données multimédia qu'il s'agisse du texte, de l'audio ou encore des images enrichi avec un système hyperliens pour relier les différentes ressources disponibles sur Internet. Depuis 1996, le W3C (*World Wide Web Consortium*) est en charge de l'évolution de ce langage qui a bien entendu été adapté en réponse à l'émergence de nouveaux besoins sur Internet aux cours des dernières années. En 2000, le HTML est devenu un standard international (ISO/IEC 15445 :2000) adopté par l'ensemble des sites web et supporté à des degrés divers par les navigateurs web (cf Fig.1.1).

FIG. 1.1 – Exemple de langage HTML (en résumé sur l'image de gauche) et le rendu dans un navigateur web (image de droite)

1.3.2.2 XML

Le XML (*Extensible Markup Language*) est un langage qui s'appuie sur des balises pour structurer des données dans une arborescence. Ce langage élaboré au sein du W3C est extensible [30]. En effet, il permet à l'utilisateur de définir ses propres balises et réutiliser celles existantes. Le XML a pour objectif de proposer un langage aussi générique que le HTML mais plus facile d'accès et à l'utilisation. Plusieurs caractéristiques présentes depuis les premières spécifications en 1998 lui ont permis d'être un langage prépondérant pour l'interopérabilité (les flux d'information sont généralement publié en RSS cf Fig.1.2) des systèmes d'information :

- Espace de noms : cette caractéristique permet de faire référence à plusieurs vocabulaires proposant des balises différentes
- Explicitation du contenu : prévoit la validation du contenu à partir d'une grammaire. Ceci ajoute des contraintes de structures qui doivent être obligatoirement respectées (la fermeture des balises est optionnelle en HTML mais nécessaire en XML)
- Support d'Unicode : Unicode est un standard qui alloue à chaque caractère d'un texte un moyen de représentation et d'identification pour la plupart des systèmes d'écriture du monde.

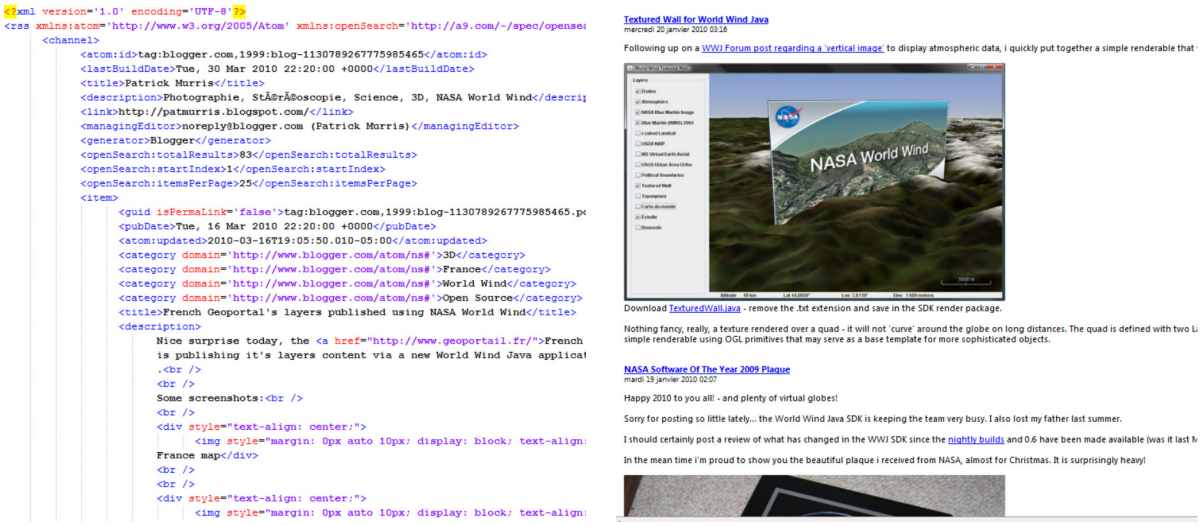


FIG. 1.2 – Exemple de langage XML (RSS tronqué sur l'image de gauche) et le rendu dans un navigateur web (tronqué sur l'image de droite)

1.3.2.3 GML

Le *Geography Markup Language* (GML) est un langage XML destiné à faciliter la manipulation et l'échange de données géographiques. La première version de ce langage [44] était définie comme une extension géographique du RDF (cf 1.4.2) dans le but de pouvoir se connecter facilement avec les nombreuses bases de données géographiques existantes. L'OGC a permis entre autre l'utilisation de propriétés d'appartenance père/fils (RDFS cf. 1.4.4) dans les dernières version du XML-schema du GML.

Le GML contient un grand nombre de primitives telles que :

- *Feature*

- *Geometry*
- *Coordinate* Reference System
- *Topology*
- *Time*
- *Dynamic feature*
- *Coverage (including geographic images)*
- *Unit of measure*
- *Directions*
- *Observations*
- *Map presentation styling rules*

En plus de ces primitives, le GML introduit la notion de *profil*. Ces profils sont des restrictions partielles de l'expressivité du GML. Ces restrictions ont été ajoutées pour faciliter l'adoption et la mise en œuvre du GML dans les applications ne nécessitant pas la manipulation de l'ensemble des primitives. Parmi les profils définis dans la norme GML, les trois profils les plus courant sont :

- *Simple Features* : apporte un support des requêtes et de transactions d'ensemble de propriétés (par exemple : WFS)
- GMJP2 : définit l'intégration de GML dans les images JPEG 2000
- Un profil pour RSS : permet de localiser géographiquement les flux d'information su le web

Bien que ces normes standardisent les processus d'échange de données, elles s'attachent surtout à structurer le contenant, pas l'information transportée. Ainsi même si le titre d'une page web peut être trouvé entre les balises $<$***TITLE***$>$ et $</$***TITLE***$>$, rien n'indique qu'il s'agisse effectivement du titre de l'article publié sur cette page. Un autre problème est celui du texte brut. Par exemple ; une description d'un article d'un flux RSS d'un journal comme lemonde.fr contient seulement le texte écrit par l'auteur. Certes, ce texte sera compréhensible directement par n'importe quelle personne mais rendre ce texte interprétable par une machine est un problème difficile. Nous emploierons le terme d'*information normalisée* pour désigner les informations non structurées contenues dans un format standardisé.

1.3.3 Extraction des entités d'intérêt

L'extraction des entités d'intérêt vise à structurer l'information normalisée pour permettre des traitements plus complexes. La complexité de ce passage de l'information non structurée à l'information structurée varie en fonction des données d'entrées et des traitements effectués[78]. La liste exhaustives des traitements possibles pour structurer l'information brute pour chaque nature de sources ouvertes dépasse le cadre de cette introduction à la représentation de l'information. Nous présenterons rapidement les points clés pour chaque nature de source et nous nous contenterons de présenter les techniques d'extraction les plus fréquemment utilisées.

1.3.3.1 Support textuel

Le support textuel constitue le principal support à partir duquel nous allons extraire des entités. En effet, lors d'une collecte d'information, la majeure partie de l'information se trouve

dans la texte contenu dans le HTML des sites web. C'est pourquoi il nous semble intéressant d'étudier le problème de mise à disposition de ces données brutes et des traitements appropriés. En effet, ces éléments sont les briques de bases à partir desquelles nous pouvons construire une représentation correcte de l'information.

Traitement Automatique du Langage Naturel
Le Traitement Automatique du Langage Naturel (TALN) est une discipline qui étudie les interactions entre les machines et les langages humains. La compréhension des langues naturelles est un problème difficile (NP-complet), cependant les avancées statistiques en apprentissage automatique ont contribué à l'amélioration des algorithmes de reconnaissance de motifs [141]. Le TALN ne se réduit pas à l'extraction d'entités mais peut s'appliquer à d'autres tâches dont voici une liste non exhaustive :

- traduction
- recherche d'information
- résumé automatique
- classification de texte
- reconnaissance de l'écriture manuscrite

Exemple Un résultat probable de l'application d'un algorithmes d'extraction d'entités nommées sur notre exemple « l'alcool bout à 176° » serait :

alcool : nom masculin, sujet de la phrase, liquide incolore et fluide qui s'obtient par distillation

bout : verbe d'état, bouillir

176° : mesure numérique contenant un nombre *176* et une unité de mesure le *degré*

Les algorithmes d'extraction sont efficaces mais dans certains cas, l'ambiguïté de la phrase peut porter à confusion. Par exemple si nous n'effectuons pas d'analyse syntaxique de la phrase avant le passage de l'extraction d'entités nommés, le mot « **bout** » aurait très bien pu être extrait comme un nom commun.

1.3.3.2 Support visuel

Le support visuel est également un support intéressant pour extraire des entités. Prenons l'exemple d'une image, voici une liste non exhaustive de traitements possibles pour extraire de l'information structurée[108] :

Détection de caractères : texte contenu dans l'image

Détection de caractéristiques d'intérêt : il est possible de détecter des objets particuliers : logos, personnes ou encore des véhicules

Extraction des données EXIF : ces données décrivent les paramètres techniques utilisés pour obtenir l'image (nom de l'appareil, du constructeur ...)

Catégorisation : lorsqu'on dispose d'un corpus d'images classées par catégories (avion, bateau, mer, ciel, montagne ...), il est possible de déterminer à quelle catégories l'image peut appartenir

Extraction des vecteurs de couleur : ce traitement est adapté à la découverte de l'environnement (intérieur/extérieur)

Toutes ces informations sont bien portées par l'image mais elle ne sont pas accessibles, la plupart du temps, sans effectuer un traitement préalable ; détection de formes, détection de visage, clustering...

1.3.3.3 Support sonore

Le support sonore est moins abondant que les deux supports précédents. Le principal traitement appliqué sur ce type de source est la transcription. La transcription substitue à chaque phonème un graphème d'une langue cible. Les méthodes de transcription qui s'appuient sur les modèles de Markov sont les plus répandus et parmi les plus efficaces [157]. De leur application résulte un nouveau support textuel contenant la même information que le support sonore. Nous pouvons alors appliquer les traitements dédiés au support textuel.

1.3.3.4 Support multimedia

La particularité du support multimedia est de rassembler les autres supports. Ainsi, nous pouvons « découper » ce support en sous-supports (texte, audio, vidéo) et leurs appliquer leurs traitements respectifs. La prise en compte des liens entre les sous-supports peut également participer à la découverte et l'extraction d'entités d'intérêt.

1.3.3.5 Support hypertexte

Dans notre contexte, le support hypertexte a la particularité de mettre en relation des ressources sur Internet. Il est intéressant d'étudier les liens entre les ressources afin d'extraire des éléments de relations entre elles. Ce système d'hyperlien permet par exemple de citer et faire référence directement la source d'un article et de la proposer au lecteur. L'extraction de ces liens se fait le plus souvent lors de la collecte d'information pour parcourir les différentes ressources sur Internet [17]. Cependant, il peut s'avérer intéressant d'étudier à nouveau ces liens dans la phase d'extraction d'information afin de structurer les relations entre ces différentes ressources.

1.3.4 Analyse des entités

L'analyse des entités extraites a pour but d'étudier ces entités structurées afin de les transformer en éléments de connaissance et les lier avec les éléments de connaissance déjà existants.

1.3.4.1 Éléments de connaissance

Les éléments de connaissance sont des éléments qui disposent de propriétés liées à leurs natures. Par exemple ; l'élément de connaissance qui représente la ville de Paris a comme propriété d'être la capitale de la France et de se situer aux coordonnées de latitude et longitude suivantes : 48.856583 2.353821. Ainsi dans la phrase « Paris se révolte », si nous associons l'entité nommée « Paris » comme une instance de l'élément de connaissance précédent, nous pouvons déterminer une nouvelle entité qui dispose à la fois du nom « Paris » et des propriétés géographiques et administratives de l'élément de connaissance.

Dans la suite de notre manuscrit nous allons nous intéresser plus particulièrement aux propriétés spatiales et temporelles. Cependant, associer des propriétés temporelles et spatiales à des entités nommées peut se révéler difficile. Cette tâche est difficile à cause de l'ambiguïté de

l'extraction de entités. Par exemple, si nous analysons l'entité « Paris » dans la phrase « Paris se révolte », Paris est personnifié dans cette phrase, ainsi l'auteur peut aussi bien vouloir représenter une ville qu'une personne. Nous sommes dans une situation similaire à celle de l'exemple de la définition du contexte (cf 1.2.1) mais il vient se greffer un problème de méronymie. En effet, ce n'est pas la ville de Paris qui se révolte mais les personnes qui manifestent leur mécontentement dans cette ville. De plus, plusieurs villes portent le nom de Paris ; il pourrait s'agir de la ville de Paris soit au Texas, soit en France. Nous nous trouvons en présence d'une nouvelle source d'ambiguïté au niveau des propriétés géographiques à associer à l'entité nommée pour en faire une instance de l'élément de connaissance.

Les éléments de connaissance varient en fonction du domaine de connaissance considéré [227]. Ainsi, les propriétés attribuées à un avion par un technicien ne seront pas les mêmes que celles attribuées par le pilote ni celles attribuées par un passager. Nous supposerons que nous disposons des moyens d'associer les entités nommées avec des instances d'éléments de connaissance tout en prenant en compte cette notion d'ambiguïté au niveau des propriétés.

1.3.4.2 Connaissance *a priori*

Lors de l'étape d'analyse, des éléments de connaissances sont déjà présents dans une base de connaissances disponible *a priori* car ils ont été déterminés en même temps que la définition du besoin de collecte. En associant les entités entre elles, cette étape va expliciter leurs relations mais également les événements d'intérêt. Cependant, l'association d'entités pour reconstruire les événements est fortement dépendant du domaine d'intérêt. En général, il est difficile pour les algorithmes d'extraction d'information d'être à la fois efficaces, précis et performants lors de ce type de traitement (cf Fig. 1.3).

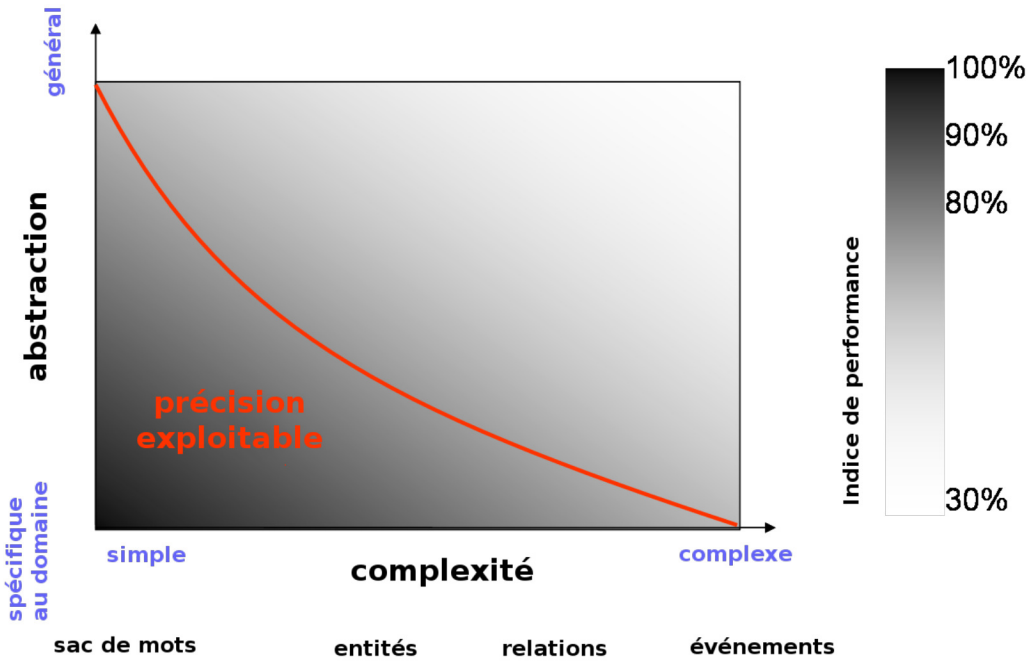


FIG. 1.3 – Comparaison de la complexité des étapes de traitement de l'information. L'abscisse représente la complexité des traitements, l'ordonnée représente la généralité du domaine d'intérêt [47].

Cette étape est primordiale pour la suite de ce manuscrit car elle va fournir les éléments de base essentiels pour la construction de notre modélisation d'événements. En effet, les événements pourront être définis grâce aux entités avec des propriétés temporelles et géographiques découvertes lors de l'extraction.

1.3.5 Indexation

L'indexation de l'information est une étape optionnelle pour la représentation de l'information. Cependant, elle peut avoir plusieurs intérêts :

l'archivage permet de retrouver des informations originales dont la source aurait disparu

la recherche fournit un moyen de parcourir le contenu de l'information par l'utilisation de filtres et de mots clés

l'amélioration de l'extraction est possible grâce à la réutilisation des informations et des entités extraites lors des traitements

l'amélioration de l'analyse se fait par l'enrichissement de la base de connaissances avec les nouveaux éléments de connaissance extraits

Ce stockage peut prendre plusieurs formes mais pour être à la fois efficace et performant, il est souvent nécessaire de stocker les informations et les entités en fonction de leurs propriétés.

1.3.5.1 Indexation du texte

L'indexation du texte va stocker les informations brutes et/ou normalisées pour fournir un moyen efficace et performant de retrouver un document à l'aide de requêtes (par mot clés, filtres, similarité). Si elle n'est pas destructrice, cette indexation va permettre d'archiver les information pour reconstruire le document d'origine. Sinon, dans la plupart des cas, seuls les termes discriminant de l'information sont gardés. En général, cette sélection est effectuée à la fois en fonction du document contenant l'information mais également du corpus de document qu'on souhaite indexer. Ce calcul se fait en comparant le poids d'un terme dans le document avec le nombre d'occurrence de ce terme dans le corpus, cette mesure statistique est appelée TF-IDF [110] (Term Frequency–Inverse Document Frequency).
Soit le document d_j et le terme t_i, alors la fréquence du terme dans le document est :

$$\mathrm{TF_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

où $n_{i,j}$ est le nombre d'occurrences du terme t_i dans d_j. Le dénominateur est le nombre d'occurrences de tous les termes dans le document d_j.

$$\mathrm{IDF_i} = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

où $|D|$ est le nombre total de documents dans le corpus et $|\{d_j : t_i \in d_j\}|$ est le nombre de documents où le terme t_i apparaît (c'est-à-dire $n_{i,j} \neq 0$).
Le poids TF-IDF se calcule ainsi :

$$\mathrm{TFIDF_{i,j}} = \mathrm{TF_{i,j}} \times \mathrm{IDF_i}$$

1.3.5.2 Indexation des images

L'indexation des images va construire un index en fonction des caractéristiques des images [137, 146]. Celles ci sont représentées par un nuage de points dans un espace de dimension élevée (l'espace de *feature*). Cette méthode d'indexation permet de définir des mesures de similarité[210] :

- similarité intra-classe : transformations admissibles entre instances d'une même classe
- similarité inter-classes : notion subjective, apprentissage

1.3.5.3 Indexation sémantique

L'indexation sémantique sert à stocker les éléments de connaissance et leurs propriétés (parfois appelées metadonnées). Les données stockées sont représentées sous formes de triples : (*Sujet*, *Prédicat*, *Objet*)[18]. Par exemple ; l'élément de connaissance qui représente la ville de Paris a comme propriété d'être la capitale de la France et de se situer aux coordonnées 48.856583, 2.353821 peut être représentée avec les cinq triples suivants :

- (**ville de Paris**, *se nomme*, Paris)
- (**ville de Paris**, *est une* , **Ville**)
- (**ville de Paris**, *est la capitale de la* , **France**)
- (**ville de Paris**, *a pour latitude*, 48.856583)
- (**ville de Paris**, *a pour longitude*, 2.353821)

où les expressions en gras sont des **éléments de connaissance**, les éléments en italique sont des *prédicats* et les expressions qui ne sont pas mises en évidence sont des littéraux. Les objets des triples peuvent soit être des littéraux, soit référencer d'autres éléments de connaissance par leur sujet.

Ce mode de représentation sert à engranger les nouvelles connaissances mais elle peut également découvrir de nouvelles connaissances si des règles d'inférence ont été définies.

Après avoir présenté les traitements de l'information, nous allons introduire les moyens de représenter les informations sémantiques issues des connaissances extraites.

1.4 Information sémantique

L'extraction d'entités est intéressante pour structurer l'information. Cependant, elle comporte des limites (cf 1.3.4.1 l'exemple de Paris les villes ou la personne) ; savoir qu'une entité nommée a été extraite ne suffit pas pour représenter complètement l'information car la sémantique présente dans l'information n'y est pas directement attachée[175]. Cette limitation vient principalement du fait que la majorité des ressources disponibles sur le web sont échangées dans un format qui ne contient pas ou contient peu d'informations sémantiques.

1.4.1 Web Sémantique

Dans le but de formaliser l'échange de ces ressources et metadonnées, le W3C a développé plusieurs langages sérialisables en XML. Nous présentons les trois principales recommandations qui visent à améliorer l'interopérabilité des ressources sur Internet pour, entre autres, faciliter les processus de traitement automatique de ces ressources.

1.4.2 RDF

Resource Description Framework (RDF) est le langage de base du Web Sémantique [134]. Il se base sur les *Unique Resource Identifier* (URI) pour référencer les ressources et les représenter à l'aide de graphes à partir d'associations de trois notions :

$$(Sujet,\ Prédicat,\ Object)$$

Sujet désigne la ressource et peut soit être anonyme, soit identifié par une URI

Prédicat désigne le type de propriété et est identifié par une URI

Objet désigne soit une ressource identifiée par une URI, soit un littéral

Ce langage est surtout utilisé pour annoter des documents et permettre un certains degré l'interopérabilité lors des échanges d'informations non structurées ; la première version de RSS (cf Fig.1.2) est basée sur le RDF.

1.4.3 SKOS

Simple Knowledge Organisation System (SKOS) est un langage basé sur le RDF dont le but est de permettre la publication de vocabulaires structurées (par exemple : thésaurus, classifications) en vue de leur réutilisation sur le Web [165].

Les représentations par concept et les propriétés de correspondance définies dans SKOS facilite la comparaison et l'alignement de concepts provenant de sources différentes. Plusieurs organisations nationales et internationales ont publié leurs vocabulaire de référence au format SKOS :

- La Bibliothèque du Congrès des États-Unis a publié le *Library of Congress Subject Headings* (LCSH) en SKOS
- La Bibliothèque Nationale de France et d'Allemagne ont publié leur vocabulaire et les correspondance entre eux et le LCSH
- Le thésaurus MeSH [213] (Medical Subject Headings) est accessible en SKOS
- L'ontologie Geonames[217] représente ses types de propriétés géographiques en SKOS

1.4.4 RDFS

RDF Schema (RDFS) est également un langage basé sur RDF dont l'objectif est de représenter des connaissances [156]. Il apporte les éléments de base pour la définition d'ontologies ou vocabulaires destinés à structurer des ressources RDF. Les principaux apports de RDFS sont la notion de *classes* et de *hiérarchie* entre ces classes pour les ressources et les notions de *domaine de définition* et *domaine d'application* pour les propriétés.

Ces principaux composants de RDFS sont également intégrés dans un langage d'ontologie plus expressif, OWL.

1.4.5 OWL

Web Ontology Language (OWL) est un langage basé sur la syntaxe RDF qui a pour but de définir des ontologies structurées [161]. Ce langage permet à la fois de décrire les concepts mais également les instances de ces concepts. En plus de reprendre certaines propriétés de RDFS (notamment la notion de hiérarchie et de domaine de définition et domaine d'application), l'OWL

définit, entre autres, de nouveaux concepts afin de faciliter la comparaison entre plusieurs ressources, classes ou propriétés.

En facilitant l'interopérabilité entre les ressources, l'OWL simplifie également l'inférence en proposant trois sous langages :

- OWL-Lite : l'expressivité est restreinte mais les algorithmes d'inférence sont efficaces
- OWL-DL : l'expressivité est meilleure qu'avec OWL-Lite mais certains algorithmes d'inférence ont une complexité exponentielle
- OWL-Full : l'expressivité est la meilleure mais il n'existe pas d'algorithmes satisfaisants pour traiter l'inférence

1.5 Conclusion

Dans ce chapitre nous avons présenté les notions de base qui nous seront utiles dans la suite de ce manuscrit :

- La représentation de l'information
- La publication de l'information disponible en sources ouvertes, notamment sur Internet
- Les problèmes d'extraction des entités contenues dans les informations pour associer des propriétés liées aux éléments de connaissance
- La représentation de la sémantique de l'information et des éléments de connaissance

Nous pouvons résumer les phases de traitement qu'à subie l'information :

1. la *source d'information* publie
2. l'*information brute (non structurée)* est normalisée
3. l'*information normalisée* est utilisée pour extraire
4. les *entités d'intérêt* sont analysées afin de mettre en évidence
5. les *relations entres entités* et les *événements*

Dans les chapitres suivants, nous rappellerons les travaux de formalisation des représentations des entités temporelles et des entités spatiales.

Chapitre 2

Représentation des entités temporelles et des entités spatiales

Sommaire

Ce chapitre présente l'intérêt de modéliser formellement les entités temporelles et les entités spatiales et dresse l'état de l'art des représentations du temps et des représentations de l'espace.

De très nombreux travaux ont été entrepris depuis que l'homme s'attache à représenter son environnement dans le temps et l'espace. Plutôt que de nous attacher à définir la nature du temps et de l'espace, nous allons nous attarder sur leur utilisation. Dans quel but représenter des entités temporelles et des entités spatiales ? Dans ce chapitre, nous avons décidé de présenter plus en détails les représentations du temps et les représentations de l'espace en nous plaçant du point de vue des entités temporelles et des entités spatiales. Cette structuration nous permet à la fois de faire le lien avec les entités du chapitre précédent et d'introduire les aspects formels des différentes représentations afin de mettre en avant les principes de raisonnements sur les connaissances temporelles et les connaissances spatiales.

2.1 Les représentations du temps

Le choix de l'unité temporelle de base a suscité de nombreuses discussions, parfois de nature philosophique, au sein de la communauté d'IA [215]. Bien que ces discussions ne convergent pas toujours, elle ont largement enrichi les connaissances du domaine, notamment en ce qui concerne l'aspect ontologique du temps : Que faut-il choisir pour faire référence au temps, le point, l'intervalle ou d'autres primitives ?

L'existence d'événements instantanés a conduit certains chercheurs à choisir le point comme unité temporelle. D'autres chercheurs ont préféré utiliser des intervalles afin de représenter de manière adéquate les assertions qui apparaissent sur une ligne du temps.

Ainsi, McDermott a proposé de représenter le temps à l'aide des points [160]. Allen [8] pense que ce choix conduit à des situations paradoxales et il opte pour l'utilisation de l'intervalle. Afin de convaincre la communauté de son choix, il illustre son point de vue en traitant un exemple de sens commun [8] : l'intervalle durant lequel une lampe est allumée rencontre l'intervalle de temps sur lequel celle-ci est éteinte. Au moment où ces deux intervalles se rencontrent, il y a un changement de la valeur de vérité de la propriété *"La lampe est allumée"*. Quelle est la valeur de vérité de cette proposition au moment de la rencontre ? Est-elle vraie, fausse ou vraie et fausse en même temps ? Lorsque les assertions sont interprétées uniquement sur des intervalles, ce type de situations est évité. En effet, il est commode d'admettre que la lampe doit être allumée ou (exclusif) éteinte à tout moment.

Dans ce qui suit, nous présentons rapidement les représentations fondées sur les instants. Ensuite, nous abordons les représentations fondées sur les intervalles.

2.1.1 Les représentations fondées sur les instants

Ce type de représentation du temps est le plus ancien et l'approche de McDermott est sans doute la plus connue [159]. Dans cette approche, l'objet de base est un *état* qui représente un instant, un moment de l'univers. L'ensemble des états est ordonné à l'aide de la relation d'ordre *"précédence ou coïncidence"*, notée : " $\leq$ " et telle que ($s_1 \leq s_2$) signifie que l'état s_1 *précède* ou *coïncide* avec l'état s_2 . Cet ensemble est également *dense*, c'est-à-dire qu'entre deux états s_1 et s_2 tels que (s_1 *précède* s_2), il existe toujours un troisième état s_3 tel que (s_1 *précède* s_3) et s_3 *précède* s_2). A chaque état de l'univers est associé un nombre réel dit **date** grâce à une fonction de datation. La structure temporelle est ainsi isomorphe à la droite des réels, d'où la non circularité du temps.

Dans l'approche de McDermott, le futur est indéterminé. En revanche, il n'y a qu'un seul passé, même s'il n'est pas connu. Ainsi, le temps est organisé en un arbre d'états représentant plusieurs futurs et un seul passé.

Sur l'ensemble des états, McDermott définit les *chroniques*. Une chronique, représentée par une branche de l'arbre des états, retrace une histoire possible du monde. Elle est représentée par un ensemble d'états convexe, totalement ordonné et non borné. Chaque état de l'univers appartient à une chronique.

Une telle représentation offre l'avantage de pouvoir définir les dates et les durées d'événements à l'aide d'un ensemble dense de points.

2.1.2 Les représentations fondées sur les intervalles

Les intervalles représentent directement des périodes de temps. Ils possèdent des durées et ne sont pas nécessairement indivisibles. Ainsi, ils constituent une abstraction des ensembles d'instants. Ces derniers peuvent constituer à leur tour une interprétation possible des intervalles. Les intervalles peuvent être classifiés selon deux types : les *intervalles convexes* et les *intervalles non convexes*.

Les intervalles convexes ont été largement considérés dans la littérature [96, 106, 159, 20, 95, 218, 8]. L'approche la plus représentative utilisant ce type d'intervalles est celle d'Allen. Un intervalle convexe représente une période de temps continue. Lorsque les intervalles convexes sont construits sur les ensembles de points [203], ils représentent alors des ensembles convexes de points à une dimension.

Un intervalle non convexe est un ensemble d'intervalles convexes dont chaque élément est appelé *composante* ou *sous-intervalle*. Ladkin [121] a introduit la notion d'*union d'intervalles convexes*, forme normale des intervalles non convexes qui permet de traiter la plupart des applications. C'est cette approche que nous considérons dans notre travail.

Dans ce qui suit, nous détaillons ces deux représentations. Nous commençons par les intervalles convexes. Ensuite, nous abordons les unions d'intervalles convexes.

2.2 Les représentation du temps fondée sur les intervalles

2.2.1 Les intervalles convexes

La représentation du temps proposée par Allen constitue le point de départ des approches fondées sur le concept d'intervalle. Allen a défini un système de 13 relations, dites *relations atomiques*, mutuellement exclusives entre les intervalles. Ce système lorsqu'il est muni de l'opération de composition et de l'opération d'addition possède une structure d'algèbre, d'où le nom d'*algèbre d'intervalles*.

Allen montre que deux intervalles quelconques peuvent se situer à l'aide de 13 relations atomiques.

D'après Allen, deux intervalles peuvent se précéder, se succéder, se chevaucher, se synchroniser au début ou à la fin, être inclus ou être égaux. Nous utilisons la terminologie de [131] pour noter ces relations :

$$E, P, M, O, S, F, D$$

pour *égal, précède, rencontre, chevauche, débute, termine, contenu*, et

$$P, M, O, S, F, D$$

pour les relations inverses *précédé, rencontré, chevauché, débuté, terminé* et *contient.*

$U = \{E, P, M, O, S, F, D, P, M, O, S, F, D\}$ est la relation universelle. L'algèbre des intervalles convexes est constituée de tous les sous-ensembles de U y compris l'ensemble vide. Il existe 2^{13} relations possibles dans l'algèbre d'intervalles. Sur l'ensemble des 2^{13} relations, Allen définit l'opération d'union ($\vee$) et l'opération d'intersection ($\wedge$). L'union de deux relations est obtenue par l'union ensembliste des relations atomiques qui représentent chacune des deux relations. Par exemple, soient $R = \{E, P, M, O\}$ et $R^{'} = \{O, S, F\}$ deux relations, nous avons :

$$R \ \vee \ R^{'} = \{E, P, M, O, S, F\}$$

L'intersection de deux relations s'obtient en appliquant l'intersection ensembliste sur les ensembles de relations atomiques qui les représentent. Sur le même exemple :

$$R \ \wedge \ R^{'} = \{O\}$$

Sur l'ensemble de ces relations, Allen définit une loi de composition interne notée o. Cette loi est associative, non commutative et possède comme élément neutre la relation d'égalité. Elle permet d'inférer de nouvelles relations à partir des relations déjà connues. Par exemple, soient R_1 et R_2 deux relations et i_1, i_2 et i_3 trois intervalles convexes. La connaissance de ($i_1 \ \ R_1 \ \ i_2$) et de ($i_2 \ \ R_2 \ \ i_3$) permet de déduire la relation $R_1 \ o \ R_2$ entre i_1 et i_3. Lorsque R_1 et R_2 sont des relations atomiques, le résultat de $R_1 \ o \ R_2$ est donné dans la table de composition d'Allen [8]. Cependant, pour deux relations quelconques, $R = \{R_1, R_2 \ldots R_n\}$ et $R^{'} = \{R_1^{'}, R_2^{'} \ldots, R_m^{'}\}$, $R \ o \ R^{'}$ est définie par :

$$R \ o \ R^{'} \ = \bigvee_{i,j} R_i \ o \ R_j^{'}$$

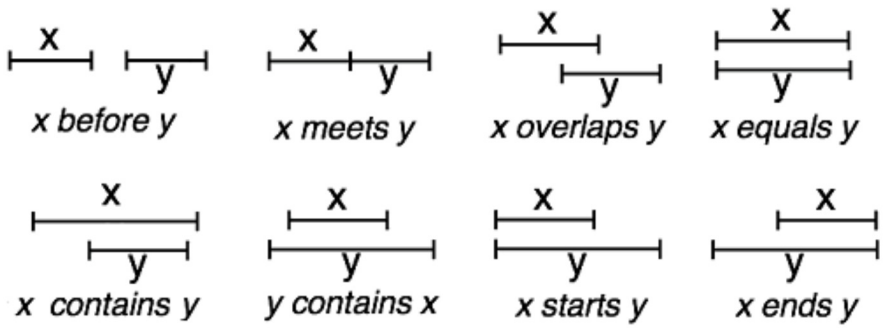


FIG. 2.1 – Relations d'Allen

Plusieurs auteurs ont étudié la théorie des intervalles convexes. Dans [215], van Benthem considère une théorie du premier ordre des intervalles convexes sur un ensemble non borné, dense et linéairement ordonné. En particulier, il a montré que cette théorie est dénombrablement catégorique, et donc décidable. Il a été montré dans [128] comment éliminer les quantificateurs d'une formule quelconque du premier ordre de cette théorie. Cette élimination est accomplie en utilisant une procédure d'élimination de quantificateurs dans la théorie des nombres rationnels [34]. Ladkin et Madux [121] proposent une reformulation algébrique de cette structure en terme des treize relations dérivables à partir de points. Hayes et Allen [5] ont formalisé l'algèbre d'Allen avec une axiomatisation où toutes les relations possibles entre les intervalles convexes

sont exprimées à l'aide de la relation *rencontre* uniquement. Cette théorie ainsi que d'autres sont présentées d'une manière détaillée dans [118].

2.2.2 Les unions d'intervalles convexes

Ces intervalles sont employés lorsque des périodes de temps discontinues sont associées à une assertion donnée. Une union d'intervalles convexes I est définie comme un ensemble d'intervalles convexes représentant chacun un sous-intervalles convexes de I. Chaque sous-intervalles convexes d'une union d'intervalles convexes appelé *maxconsubint* [123] est maximal. En d'autres termes, un maxconsubint z de I est un sous-intervalle convexe de I tel que :

$$\forall j : j \textit{ est un sous-intervalle convexe de } I \Rightarrow \neg(j\ M \vee M \vee O \vee O \vee S \vee D \vee F\ z)$$

Ce qui signifie que tous les sous-intervalles convexes de I qui sont *disjoints* de z doivent précéder (ou succéder) z et ne doivent pas rencontrer (ou être rencontré par) z, c'est-à-dire :

$$\text{pour tout sous-intervalle convexe } j \text{ de } I, \quad j\ P \vee P\ z$$

$$\text{pour tout sous-intervalle convexe } jdeI, \quad j \text{ non } (M \vee M)\ z$$

Dans [121], Ladkin présente une taxonomie de relations entre les unions d'intervalles convexes. Il démontre que le nombre de relations entre les unions d'intervalles convexes est au moins exponentiel en nombre de leurs *maxconsubints*. Afin d'éviter une explosion combinatoire, il définit des relations de base entre les unions d'intervalles convexes ne dépendant pas des nombres de leurs composantes. Ces relations sont considérées comme une généralisation des relations définies par Allen. Elles peuvent être classifiées en deux catégories. Les relations de la première catégorie sont obtenues en appliquant des *constructeurs de relations* sur les relations d'Allen. Ces constructeurs sont : *Mostly, Always, Partially, Sometimes* et *Disjunction*. Plus précisément, soient I et $I^{'}$ deux unions d'intervalles convexes et, soit R une relation d'Allen :

- I *Mostly* R $I^{'}$ si et seulement si pour toute composante $i^{'}$ de $I^{'}$, il existe une composante i de I telles que $i\ \ R\ \ i^{'}$,
- I *Always* R $I^{'}$ si et seulement si I *Mostly* R $I^{'}$ et $I^{'}$ *Mostly* $R\ \ I$,
- I *Sometimes* R $I^{'}$ si et seulement si il existe au moins une paire d'intervalles i et $i^{'}$ appartenant respectivement à I et $I^{'}$ et tels que $i\ \ R\ \ i^{'}$,
- I *Partially* R $I^{'}$ si et seulement si certaines composantes de I et de $I^{'}$ sont liées à l'aide de R alors que les autres composantes sont disjointes.
- I *Disjunction* R $I^{'}$ si et seulement si pour toute paire d'intervalles i et $i^{'}$ appartenant respectivement à I et $I^{'}$, nous avons $i\ \ R\ \ i^{'}$ ou $i\ \ P\ \ i^{'}$ ou $i\ \ P\ \ i^{'}$.

Les relations de la seconde catégorie sont celles définies directement sur les unions d'intervalles convexes. Par exemple la relation *précède* lie deux unions d'intervalles convexes I et $I^{'}$ si et seulement si toutes les composantes de I précèdent toutes les composantes de $I^{'}$. Enfin, Ladkin introduit une nouvelle relation symétrique qu'il nomme *bars*. Cette relation lie deux unions d'intervalles convexes si et seulement si leur union est un intervalle convexe.

Contrairement à l'algèbre des relations entre les intervalles convexes, l'algèbre des relations non convexes est infinie et ne permet pas de définir une classe d'intervalles à n sous-intervalles, pour chaque entier n. Dans [123], le même auteur propose une extension des intervalles d'Allen permettant de prendre en compte les unions d'intervalles convexes. Il propose de nouvelles

primitives destinées à instancier des unions d'intervalles convexes sous une forme permettant de représenter les unités de temps telles que l'année, le mois, le jour, *etc.*

D'autres auteurs se sont intéressés à ce type d'intervalles. Ligozat, par exemple, a proposé une autre représentation [144]. Celle-ci consiste à exprimer une union d'intervalles convexes à l'aide d'une séquence de points, finie et ordonnée linéairement. Plus précisément, une union d'intervalles convexes est exprimée à l'aide de la séquence :

$$\langle a_1, \ldots, a_{2n} \rangle$$

telle que pour tout i, $1 \leq i \leq (2n-1)$, $a_i < a_{i+1}$. Une composante de l'intervalle est, par conséquent, un couple de points $\langle a_i, a_{i+1} \rangle$ où i est un entier impair.

Les relations entre les intervalles généralisés sont exprimées sous forme de conjonctions de conditions sur les points de la séquence. Morris, Shoaff et Khatib [170] ont proposé une étude de la notion des relations non-convexes introduites par Ladkin.

2.3 Les représentations de l'espace

La sélection d'une représentation de l'espace va impacter le traitement logique et algorithmique du problème de représentation des entités spatiales. La géométrie met à notre disposition plusieurs modèles mathématiques afin de d'obtenir une représentation satisfaisante en fonction des contraintes induites par les entités spatiales. La géométrie des mathématiciens et les géométrie du monde sensible sont les deux modèles mathématiques les plus connus que nous allons détailler.

2.3.1 La géométrie des mathématiciens

Les premières logiques de l'espace sont apparues avec les observations de l'espace par Euclide. Son système d'axiomes de la géométrie est à la base de l'analyse logique de notre perception de l'espace. Ce système d'axiomes géométrique a entraîné la mise en place d'un système de déduction permettant de démontrer de nouveaux prédicats à partir de prédicats vrais. Ce n'est qu'à partir de l'invention des géométries non euclidiennes que les mathématiciens étudièrent sérieusement les fondements logique de la géométrie[14].

Les géométries absolue, euclidienne et hyperbolique sont construites à partir d'un principe fondamental ; le point est le concept théorique désignant la plus petite parcelle concevable d'espace. Dans cet espace, le calcul des prédicats du premier ordre est utilisé pour décrire les relations entre les variables qui représentent des points. Une relation entre trois points (x, y, z) peut être par exemple « x est plus éloigné de y que de z ». Cependant, les modèles de la géométrie peuvent varier selon les auteurs. Ainsi Szmielew[207] et Tarski[208] vont choisir les prédicats $B(x,y,z)$: « y est situé entre x et z » et $D(x,y,z,u)$: « x est aussi éloigné de y que z l'est de u » pour décrire une structure relationnelle de la forme (W, B, D) où W contient l'ensemble des points qui sont éléments des relations des prédicats B et D. Robison [191] énonce le prédicat $J(x,y,z)$: « x est plus près de y que de z » pour construire une géométrie à partir de la structure relationnelle suivante (W,J) où W contient l'ensemble des points qui sont éléments des relations du prédicat J.

L'étude des modèles des géométries ainsi obtenues permet de définir la représentabilité, la complétude et la décidabilité des systèmes d'axiomes correspondants.

2.3.2 La géométrie du monde sensible

La géométrie du monde sensible vise à prendre en compte les connaissance partagée par tous lors du raisonnement sur le monde. Dans [99], Hayes essaye de formaliser l'ensemble des connaissances intuitives partagées sur le monde. Cette approche qui pose la question de la sélection des éléments de base communs se rapproche des problèmes étudiés avec les ontologies. Quels sont les élémens et les relations communes aux représentations ? Plutôt que de répondre directement par des mesures quantitatives, la plupart des travaux dans ce domaine font intervenir les relations qualitatives afin de représenter les connaissances. En effet, dans certains scénario, les données quantitatives posent parfois des problèmes :

- Le coût en ressources est trop élevé pour traiter les données
- Les données quantitatives ne sont pas forcément indépendantes entre elles
- La surcharge de données numériques peut empêcher l'utilisateur de distinguer une information intéressante

Le raisonnement spatial qualitatif est un moyen de représentation des connaissances afin de faciliter la manipulation des informations spatiales qualitatives à l'intérieur de modèles formels.

2.4 L'analyse spatiale qualitative

D'après [72], l'analyse spatiale qualitative permet de catégoriser une infinité de représentations spatiales possibles en un nombre fini de configurations. Chacune de ces configurations va être déterminée par un ensemble de concepts spatiaux communs à un domaine. Cependant, cette restriction par concepts peut entraîner des différence d'interprétations si le créateur et l'utilisateur ne partagent pas la même définition des concepts employés.

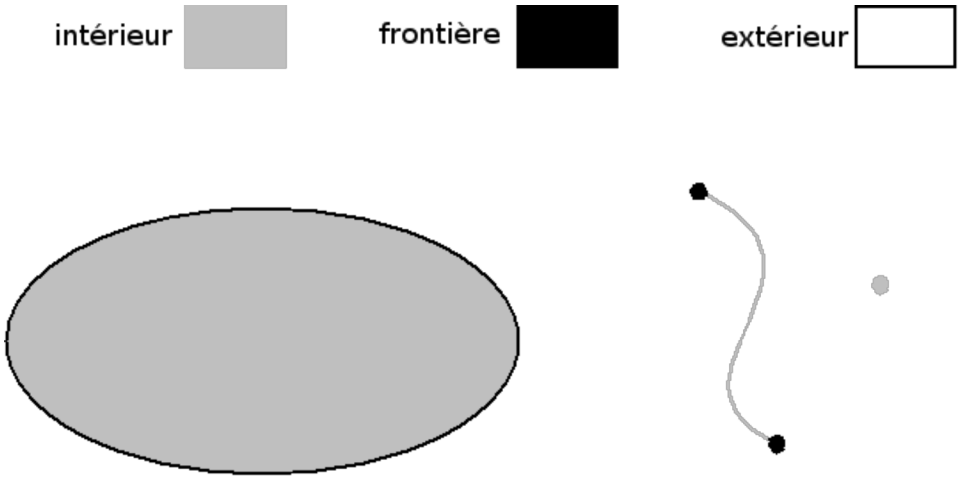


FIG. 2.2 – Représentation de l'intérieur, de la frontière et de l'extérieur pour la région, la ligne et le point.

2.4.1 La topologie

Les définitions topologiques fondamentales définies par Worboys [226] dans un contexte spatial sont :

- Soit S un espace topologique qui a un ensemble de voisinages associé. Soit X un sous-ensemble de points de S et x un point individuel de S. x est proche de X si chaque voisinage de x contient des points de X
- La fermeture de X est l'union de l'espace topologique formé par tous les points proches de X, elle est notées $\bar{X}$
- Le complémentaire X' de X est l'ensemble des points n'appartenant pas à X
- L'intérieur de X représente tous les points appartenant à X et n'étant pas proches de X' le complémentaire de X. Il est noté $\overset{\circ}{X}$
- La frontière de X représente tous les points proches de X et de X', elle est notée∂X
- L'extérieur de X représente le complémentaire de la fermeture de X et elle est notée X^{-}

Les relations topologiques (Fig.2.3) entre deux surfaces X et Y sont :

- X *touche* Y existe une partie de la frontière de X commune à la frontière de Y, et si X est à l'intérieur de Y
- X *recouvre partiellement* Y si et seulement si il existe une partie de l'intérieur de X commune à l'intérieur de Y, et si une partie de l'intérieur de X est commune à l'extérieur de Y
- X *couvre* Y si et seulement si Y est un sous-ensemble de X (Yest donc à l'intérieur de X) et s'il existe une partie de la frontière de X commune à la frontière de Y
- X *est dans* Y si et seulement si X est un sous-ensemble de Y, et si il n'existe aucune partie de la frontière de X commune à la frontière de Y

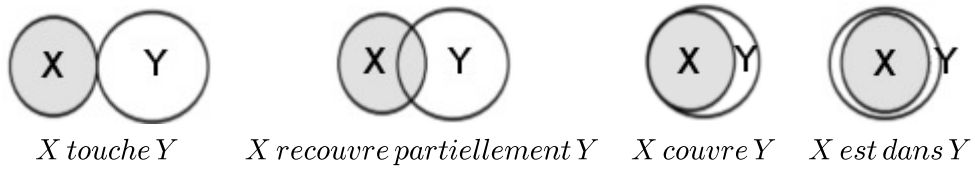


FIG. 2.3 – Relations topologiques entre deux régions X et Y

Nous allons présenter les deux modèles d'analyse spatiale qualitative topologique les plus fréquents :

2.4.1.1 Le modèle RCC-8

RCC-8 [39] décrit l'ensemble des configurations possibles pour deux régions dans un espace à deux dimensions. Ce modèle est une interprétation dans l'espace de relations d'Allen, il rassemble ces relations en deux dimensions pour représenter les transitions entre deux entités spatiales [83].

Les relations spatiales définies dans le modèle RCC-8 sont les suivantes : *Disconnected* (DC), *Externally Connected* (EC), *Partially Overlapping* (PO), *Equal* (EQ), *Tangential Proper Part* (TPP), *Tangential Proper Part inverse* (TPPi), *Non-Tangential Proper Part* (NTPP), *Non-Tangential Proper Part inverse* (NTPPi).

2.4.1.2 Le modèle des intersections

Egenhofer [76, 74] a défini formellement un modèle d'intersection pour représenter les relations entre des régions sans trou :

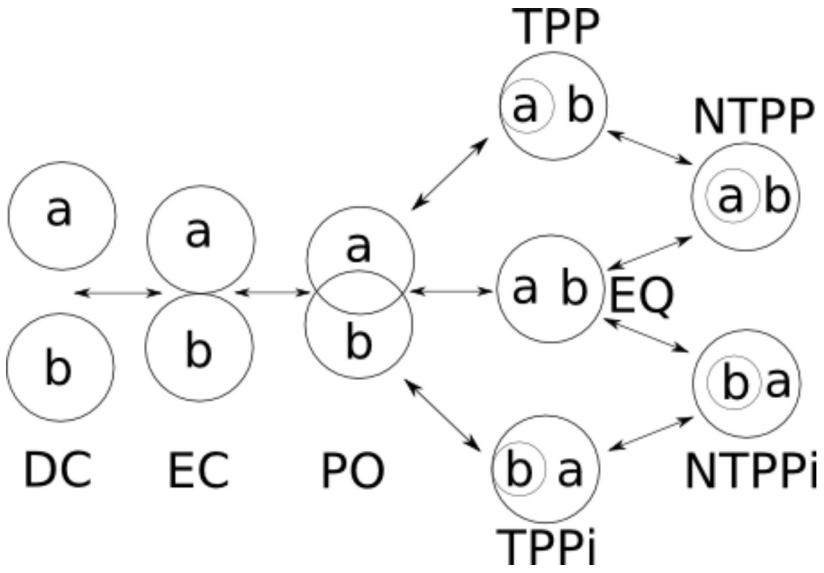


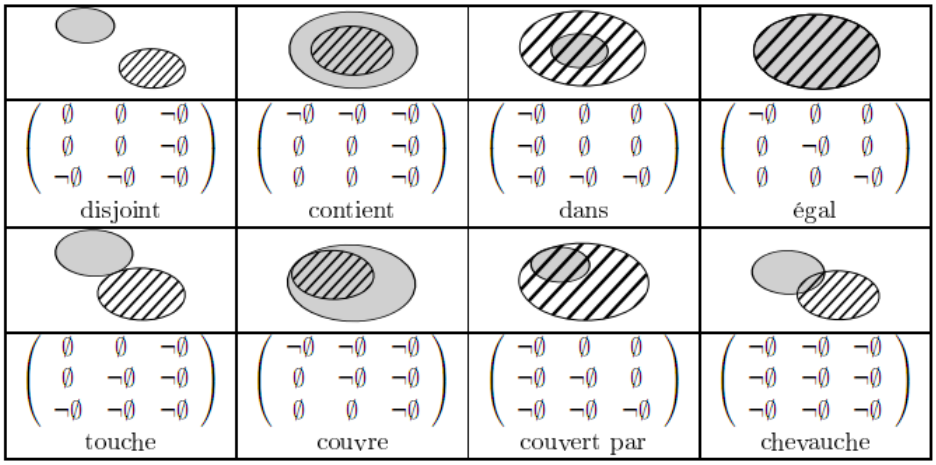
FIG. 2.4 – Représentation des relations RCC-8 et les liens de voisinage

- L'intérieur d'une région Y est définie par l'union des tous les ouverts qui sont contenus dans Y, i.e. l'intérieur de Y est le plus grand ouvert contenu dans Y.
- Le complémentaire d'une région Y dans l'espace $\mathbb{R}^2$ ou l'extérieur d'une région Y est l'ensemble des points de $\mathbb{R}^2$ ne contenant pas Y
- La fermeture d'une région Y est définie par l'intersection de tous les fermés contenant Y, i.e. la fermeture de Y est le plus petit ensemble fermé contenant Y
- La frontière d'une région Y est le résultat de l'intersection de la fermeture de Y et de la fermeture du complémentaire de Y. La frontière est un ensemble fermé.

Il existe neuf relations possibles entre l'intérieur, l'extérieur et la frontière de chaque ensemble de points X et Y. Ces intersections sont représentées sous la forme de la matrice suivante où les valeurs correspondent au résultat des intersections (avec $\emptyset$ désigne l'ensemble vide et $\neg\emptyset$ désigne l'ensemble non vide) :

$$\begin{pmatrix} \overset{\circ}{A} \cap \overset{\circ}{B} & \overset{\circ}{A} \cap \partial B & \overset{\circ}{A} \cap B^{-} \\ \partial A \cap \overset{\circ}{B} & \partial A \cap \partial B & \partial A \cap B^{-} \\ A^{-} \cap \overset{\circ}{B} & A^{-} \cap \partial B & A^{-} \cap B^{-} \end{pmatrix}$$

Les relations possibles entre deux régions sans trous dans un espace à deux dimensions sont présentées dans le tableau 2.1.



TAB. 2.1 – Relations topologiques entre deux régions sans trous dans un espace à deux dimensions

2.5 Ontologies Spatio-temporelles

Les ontologies spatio-temporelles vont nous intéresser dans la mesure où elles vont nous permettre de formaliser les entités temporelles et spatiales découvertes dans l'information que nous aurons récupérée. La théorie de l'information ne peut représenter les entités spatio-temporelles que de manière limitée [43]. Les ontologies spatio-temporelles se déclinent en deux approches [92] :

2.5.1 Ontologie de type SNAP

Une ontologie de type SNAP est photographie instantanée (*snapshot*) des entités qui existent à un instant donné. Ainsi chaque ontologie SNAP va décrire les entités et leurs relations sans prendre en compte le temps. La représentation de l'évolution des entités dans le temps fait intervenir un index temporel qui « ordonne » (en fonction de la nature du temps) chaque SNAP.

2.5.2 Ontologie de type SPAN

Une ontologie de type SPAN définit l'évolution des entités géographiques en s'attachant à définir chaque intervalle de temps durant lequel l'entité existe. Une ontologie de type SPAN contient l'ensemble de la « vie » (passé, présent et futur) de l'entité en question.

Les entités décrites dans une ontologie de type SNAP ont leur homologue dans une ontologie de type SPAN mais ils ne sont pas identiques car ils ne déterminent pas les même propriétés dans l'espace et le temps.

2.6 Conclusion

Ce chapitre a répertorié des techniques de représentation des entités temporelles et des entités spatiales et il introduit les formalismes de modélisation des relations entre de telles entités. Nous disposons maintenant des outils nécessaires à la construction d'un modèle temporel, spatial et sémantique de représentation des événements.

Deuxième partie

Contributions : Modèle temporel, spatial et sémantique pour la découverte de relations entre événements

Chapitre 3

Modèle temporel, spatial et sémantique

Sommaire

Dans ce chapitre, nous rappelons l'intérêt d'un modèle formel de représentation des événements afin de découvrir de nouvelles relations. Nous introduisons une représentation temporelle et une représentation spatiale des entités qui définissent la notion d'événement dans notre modèle en précisant les apports sémantiques. Cette notion constitue le socle de la modélisation que nous présentons. Nous étudions ensuite l'évolution de ces événements en définissant la notion de phénomène.

3.1 Motivations

Nous cherchons à obtenir une modélisation pertinente de l'information, pour cela nous devons être capable de représenter correctement les descriptions temporelles et spatiales des entités qui les composent. Nous cherchons également à retrouver les liens entre différentes informations grâce à la modélisation des relations de connexion entre ces entités. Pour cela, nous devons étudier les relations temporelles et spatiales entre les entités extraites. Généralement, un intervalle est un choix convenable pour représenter une date (Vendredi 10 juillet 2009) ou une période (dans les années 60) mais la composante de temps décrite dans « Au début du mois » devient difficile à représenter ainsi.

La modélisation des événements garantie une représentation temporelle et spatiale efficace des systèmes complexes [93, 226]. Nous introduisons des méthodes de raisonnement pour découvrir de nouvelles relations entre les événements.

Nous cherchons à modéliser de façon formelle l'information, ou plutôt les entités temporelles et entités spatiales qu'elle contient, en ensemble d'événements et de relations entre ces événements. Cette modélisation a de multiples objectifs :

- Formaliser la description des événements en s'appuyant sur les aspects temporels, spatiaux et sémantiques
- Proposer une modèle formel de représentation de ces événements et de leurs relations
- Introduire les propriétés définissant comment regrouper ces événements en fonction de leur caractéristiques
- Permettre de raisonner sur ces événements afin de découvrir de nouvelles relations non évidentes *a priori*

3.2 Représentation des entités temporelles

Comme nous l'avons vu au chapitre précédent, les entités décrites en langage naturel sont parfois ambiguës. La représentation de l'incertitude temporelle est une solution afin de ne pas négliger cette ambiguïté en la formalisant telle une partie essentielle de la représentation temporelle.

Dans la suite de ce manuscrit, nous considérons que le temps est divisé en unités sans durée appelées « instants ». Nous supposons que le temps est composé d'un ensemble d'instants ordonnés selon une relation d'ordre strict $<$ qui dispose de ces propriétés :

Linéarité : $\forall t, \forall s,\ t < s \vee s < t \vee s = t$

Densité : $\forall s, \forall e, \exists t$ tels que $s < e \rightarrow s < t < e$

Continuité : $\forall t, \exists i, \exists j$ tels que $i < t \wedge t < j$

où e, i, j, s, t sont des instants.

3.2.1 Intervalle flou

Un intervalle flou (noté I) décrit une entité temporelle en respectant l'ambiguïté présente dans le texte original.

Définition 3.2.1 *Intervalle Flou*

$$I = [s, e] \text{ tel que } \forall t \in I,\ s_{min} \leq t \leq e_{max}$$

avec $s_{min} \leq s \leq s_{max}$ *et* $e_{min} \leq e \leq e_{max}$ *où* $s_{min}, s_{max}, e_{min}, e_{max}, e, s, t$ *sont des instants. Ainsi, la structure temporelle de notre modèle est assez flexible pour décrire les entités temporelles.*

3.2.2 Relation entre intervalles temporels flous

Pour modéliser des expressions temporelles, nous introduisons des relations temporelles entre deux intervalles temporels flous :

Définition 3.2.2 *Précède*
Un intervalle temporel flou $I_1 = [s_1, e_1]$ *précède (respectivement précède strictement) un intervalle temporel flou* $I_2 = [s_2, e_2]$ *noté* $I_1 \leq_I I_2$ *(resp.* $I_1 <_I I_2$*) si et seulement si* $e_1 \leq s_2$ *(resp.* $e_1 < s_2$*).*

Définition 3.2.3 *Succède*
Un intervalle temporel flou $I_1 = [s_1, e_1]$ *succède (respectivement succède strictement) un intervalle temporel flou* $I_2 = [s_2, e_2]$ *noté* $I_1 \geq_I I_2$ *(resp.* $I_1 >_I I_2$*) si et seulement si* $s_1 \geq e_2$ *(resp.* $s_1 > e_2$*).*

3.3 Représentation des entités spatiales

Pour représenter les entités spatiales, nous allons nous appuyer sur les définitions d'entités spatiales introduites par [138] :

Définition 3.3.1 *Entité Spatiale (SP)*
- *Entité Spatiale Absolue : peut être directement localisée sur une carte.*
- *Entité Spatiale Relative : composée d'au moins une entité spatiale absolue et de relations topologiques.*

Ainsi, une entité spatiale relative est définie récursivement par d'autres entités spatiales. La région spatiale décrite par une entité spatiale dépend de l'entité spatiale absolue, des extensions de cette région et des relations topologiques.

Exemple L'entité spatiale « Près de Caen » est une entité spatiale relative composée d'une entité spatiale absolue « Caen » et d'une relation d'adjacence « Près de ».

3.4 Relations de connexion temporelle et spatiale

3.4.1 Connexion temporelle

Définition 3.4.1 *Connexion Temporelle*
Deux intervalles I_1,I_2 sont connectés temporellement si et seulement si ils vérifient une des relations d'Allen suivantes : chevauche, égal, contient, commence, touche et leurs inverses. Nous notons cette relation $t-connected(I_1, I_2)$.

Par exemple, la relation entre les deux intervalles temporels flous « *during May* » (I_1) et « *at the end of April* » (I_2) peut être représenté de la manière suivante : On a $I_1 = [s_1, e_1]$ et $I_2 = [s_2, e_2]$ où

- 29 $Avril$ 2008 $< s_1 <$ 2 Mai 2008
- 30 Mai 2008 $< e_1 <$ 2 $Juin$ 2008
- 17 $Avril$ 2008 $< s_2 <$ 21 $Avril$ 2008
- 29 $Avril$ 2008 $< e_2 <$ 2 Mai 2008

La relation $t-connected(I_1, I_2)$ est satisfaite car il existe un chevauchement entre I_1 et I_2. Une représentation par intervalles simples n'aurait pas permis de la trouver.

3.4.2 Connexion spatiale

Nous introduisons également une relation de connexion entre deux entités spatiales :

Définition 3.4.2 *Connexion Spatiale*
Deux entités spatiales SP_1, SP_2 sont connectées spatialement si et seulement si elles vérifient une des relations RCC-8 suivantes (Figure 2.4) : Externally Connected *(EC),* Partially Overlapping *(PO),* Equal *(EQ),* Tangential Proper Part *(TPP),* Tangential Proper Part inverse *(TPPi),* Non-Tangential Proper Part *(NTPP),* Non-Tangential Proper Part inverse *(NTPPi).*
On notera cette relation $sp-connected(SP_1, SP_2)$.

Supposons que l'on reçoive une alerte décrivant deux entités spatiales « *the earthquake in Sichuan* » (SP_1) et « *destruction near Chengdu* » (SP_2). Chengdu est la capitale de la province du Sichuan, il existe une relation de type NTTPi entre SP_1 et SP_2. La relation de connexion $sp-connected(SP_1, SP_2)$ est donc satisfaite. Dans la section suivante, nous proposons d'étendre la modélisation reposant sur les entités temporelles et spatiales grâce à la description structurée des propriétés sémantiques.

3.5 Événement

Dans notre modèle, un événement est constitué d'entités. Au minimum une entité temporelle et une entité spatiale définissent une description figée de ce qui se passe à un endroit durant un intervalle de temps, cette représentation des événements est courante [36, 59, 84] mais comporte des limitations, notamment en terme de description sémantique de l'événement. Nous proposons d'enrichir cette représentation des événements en introduisant des propriétés sémantiques. Notre but est de regrouper les événements qui partagent des significations sémantiques similaires pour en déduire de nouveaux liens entre ces événements.

Définition 3.5.1 *Evénement*

Un événement E est composé d'un intervalle temporel I, d'une entité spatiale SP et d'une propriété sémantique S noté $E\langle I, SP, S\rangle$.

3.5.1 Propriété Sémantique

L'intérêt d'une association des propriétés sémantiques et d'une modélisation temporelle et spatiale des événements est qu'elles se complètent. En effet ; les propriétés sémantiques permettent de découvrir de nouvelles relations mais selon le contexte deux événements avec les mêmes propriétés sémantiques peuvent se décrire avec des entités spatiales et temporelles différentes. Nous avons défini la propriété sémantique S d'un événement à partir d'une structure hiérarchique issue d'une ontologie. Deux propriétés sémantiques *partagent une similarité sémantique* si et seulement si ces propriétés sémantiques sont les mêmes ou s'il existe une relation hiérarchique entre elles.

Exemple

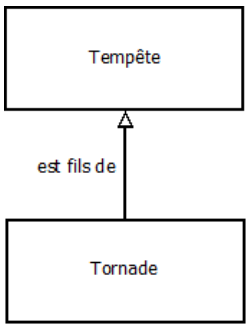


FIG. 3.1 – Relation hiérarchique entre les propriétés sémantiques *Tempête* et *Tornade*.

Une propriété sémantique *Tempête* partage une similarité sémantique avec une propriété sémantique *Tornade* car une *Tornade* est un fils de *Tempête* d'après la relation de hiérarchie de la figure 3.1.

3.5.2 Relation de similarité entre deux événements

Une relation de similarité entre deux événements indique l'existence d'un chemin constitué de propriétés de relations hiérarchiques entre ces deux entités qui ne passe pas par la racine de l'ontologie de événements.

Supposons l'ontologie des événements Fig.3.2.

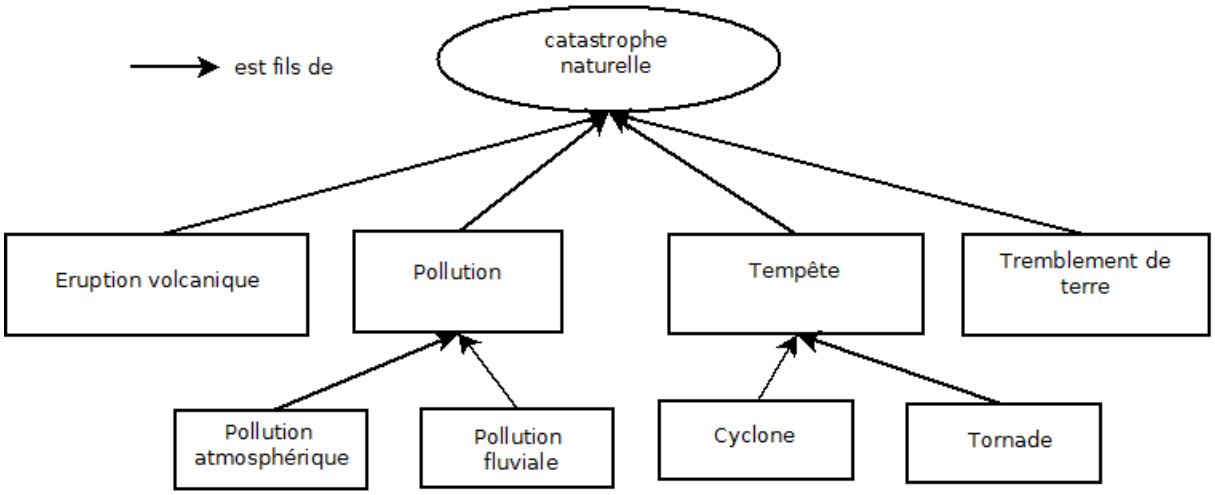


FIG. 3.2 – Exemple d'ontologie des événements avec des relations de filiation

Dans cette ontologie, il existe les relations de similarité suivantes :

- *Pollution* et *Pollution atmosphérique*
- *Pollution* et *Pollution fluviale*
- *Pollution atmosphérique et Pollution fluviale*
- *Tempête et Cyclone*
- *Tempête et Tornade*
- *Cyclone et Tornade*

3.6 Phénomène

Un phénomène est un ensemble d'événements dont les propriétés sémantiques partagent une similarité sémantique avec une propriété commune S. Dans notre modèle, un phénomène est une représentation d'une entité sémantique qui évolue temporellement et spatialement. Ainsi, un événement est une description statique et partielle d'un phénomène. Ce dernier est l'union de tous les événements et des relations entre eux. Un phénomène est noté $Ph\langle S\rangle = \cup E\langle I_E, SP_E, S_E\rangle$ où S est une propriété sémantique et $E\langle I_E, SP_E, S_E\rangle$ sont des événements dont les propriétés sémantiques S_E partagent une similarité sémantique avec S.

Exemple Une tempête (un phénomène Ph avec comme propriété sémantique Tempête défini dans une ontologie des catastrophes naturelles) a été détectée au large (premier événement E_1) d'un port de commerce, il s'est transformé en cyclone dévastant les infrastructures à l'intérieur des terres (second événement E_2). Cette alerte est représentée par deux événements et un phénomène (où $I_1 <_I I_2$) : $E_1\langle I_1, Offshore, Tempest\rangle$, $E_2\langle I_2, Inland, Cyclone\rangle$, $Ph\langle Tempest\rangle = \cup\{E_1, E_2\}$.

Un phénomène apparaît au moment où un événement est découvert. Dans la suite, nous discuterons des évolutions temporelles et spatiales d'un phénomène et des conséquences vis à vis de ses événements.

3.6.1 Évolution Temporelle : Occurrence

L'occurrence décrit l'évolution dans le temps d'un phénomène. Ce comportement définit la fréquence d'apparition d'un événement associé au phénomène. Nous proposons trois occurrences : **continuité**, **périodicité**, **aléatoire**.

3.6.1.1 Continuité

La continuité est une forme d'évolution temporelle constante au cours du temps (Fig.3.3). Elle se décompose en trois phases :

- Apparition du phénomène
- Validité continue du phénomène durant un intervalle de temps
- Disparition du phénomène

Cette évolution est unique. C'est-à-dire que le phénomène n'existe que pendant que un intervalle de temps valide sans pouvoir réapparaître ni dans le futur ni dans le passé. Cependant cette évolution temporelle peut se poursuivre dans le futur ou s'étendre dans le passé.

L'évolution temporelle de type continuité est notée : Co.

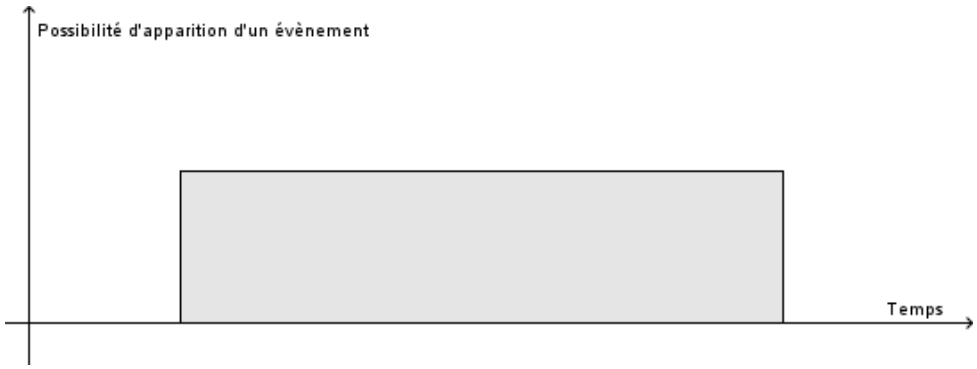


FIG. 3.3 – Exemple de représentation de la propriété d'occurrence de type **continuité** au cours du temps

3.6.1.2 Périodicité

La périodicité est une forme d'évolution temporelle répétée après un délai constant (Fig.3.4). Un comportement en trois phases se reproduit régulièrement :

- Apparition du phénomène
- Validité du phénomène durant un cours intervalle de temps
- Disparition du phénomène

Le phénomène apparaît et disparaît à intervalles réguliers un nombre indéfini de fois. L'évolution temporelle va s'étendre dans le passé et se poursuivre dans le futur sous la forme de l'apparition antérieure ou postérieure du même comportement avec la même régularité. Contrairement à la continuité, les intervalles de temps durant lesquels le phénomène est valide ne peuvent pas être prolongés.

L'évolution temporelle de type continuité est notée : Pe.

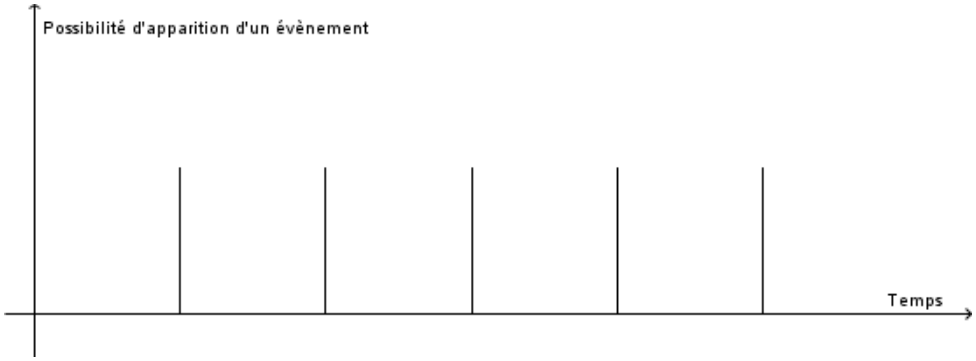


FIG. 3.4 – Exemple de représentation de la propriété d'occurrence de type **périodicité** au cours du temps

3.6.1.3 Aléatoire

L'aléatoire est une forme d'évolution temporelle répétée sans régularité au cours du temps (Fig.3.5). Un comportement en trois phases se reproduit régulièrement :

- Apparition du phénomène
- Validité du phénomène durant un intervalle de temps de durée variable
- Disparition du phénomène

Le phénomène apparaît et disparaît à intervalles non réguliers un nombre indéfini de fois et chaque période de validité est définie sur un intervalle de temps indépendant des autres. L'évolution temporelle va s'étendre dans le passé et se poursuivre dans le futur de manière irrégulière et à validité changeante.

L'évolution temporelle de type continuité est notée : *Al*.

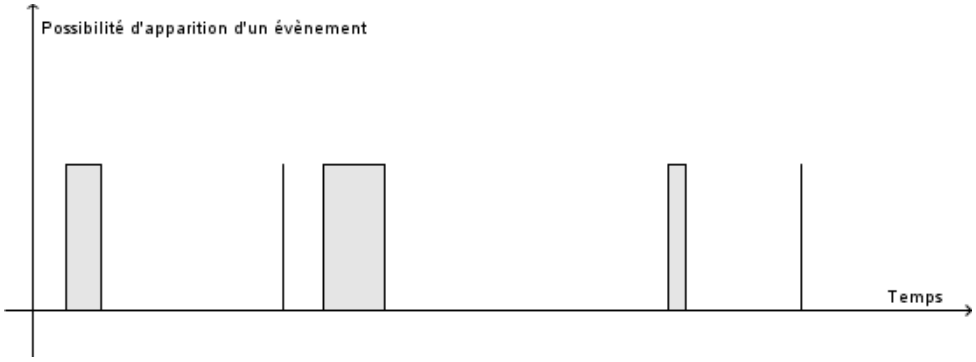


FIG. 3.5 – Exemple de représentation de la propriété d'occurrence de type **aléatoire** au cours du temps

3.6.1.4 Propriété d'occurrence temporelle

Nous supposons que, pour chaque domaine d'application, au moins une de ces occurrences a été assignée pour chaque type de phénomène. Un phénomène Ph avec une propriété sémantique S et son occurrence O sera noté $Ph\langle O, S\rangle$.

La *périodicité* et l'*aléatoire* définissent la même propriété d'union des intervalles temporels :

Soit $O \in \{Pe, Al\}$
$\forall E\langle I_E, SP_E, S_E\rangle \in Ph\langle O, S\rangle$, où $O = \cup\{I_E\}$ est une union d'intervalles convexes.
La *continuité* a la propriété d'union d'intervalles temporels suivante : Soit $O = Co$

$$\begin{gathered}\forall E\langle I_E, SP_E, S_E\rangle \in Ph\langle Co, S\rangle, \forall t \text{ un instant } \in I, \\ \exists i_{min}, \exists i_{max} \text{ des instants tels que} \\ Co = [i_{min}, i_{max}] \text{ et } i_{min} \leq t \wedge t \leq i_{max}\end{gathered}$$

Nous pouvons en déduire une première propriété générale :

Propriété 3.6.1

Soit $E\langle I_E, SP_E, S_E\rangle$ un événement et $Ph\langle O, S\rangle$ un phénomène où $O \in \{Co, Pe, Al\}$,

$$si\ E\langle I_E, SP_E, S_E\rangle \in Ph\langle O, S\rangle$$

$$alors\ on\ a\ t{-}connected(I_E, I_O) \tag{3.1}$$

où I_O est l'union des intervalles temporels inclus dans O.

3.6.2 Évolution Spatiale : Transformation

Un phénomène évolue au cours du temps et dans l'espace. La plupart du temps, cette évolution va engendrer l'apparition de nouveaux événements dans les régions affectées par le phénomène. Nous avons choisi de décomposer l'évolution spatiale (notée T) d'un phénomène en trois transformations spatiales de base : **translation**, **rotation**, **homothétie.**

3.6.2.1 Translation

La translation est une forme d'évolution spatiale suivant un vecteur (Fig.3.6). Elle se décompose en trois phases :

- Apparition du phénomène
- Validité du phénomène sur la région de l'espace décrite par le vecteur de la translation
- Disparition du phénomène

Cependant, cette évolution spatiale peut être précédée ou suivie d'autres transformations qui définiront son passé et son futur.

L'évolution spatiale de type translation est notée : Tr.

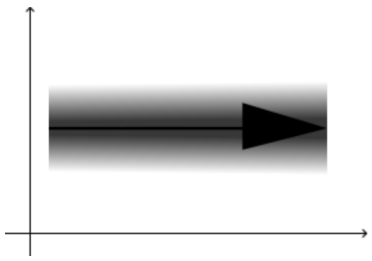


FIG. 3.6 – Exemple de représentation de la propriété de transformation de type **translation** dans l'espace

3.6.2.2 Rotation

La rotation est une forme d'évolution spatiale suivant un arc de cercle(Fig.3.7). Elle se décompose en trois phases :

- Apparition du phénomène
- Validité du phénomène sur la région de l'espace décrite par l'arc de cercle de la rotation
- Disparition du phénomène

Cependant, cette évolution spatiale peut être précédée ou suivie d'autres transformations qui définiront son passé et son futur.

L'évolution spatiale de type translation est notée : Ro.

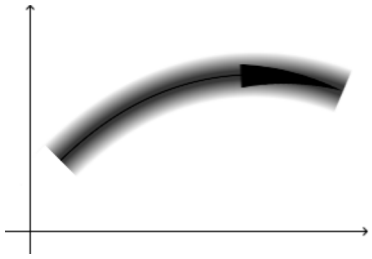


FIG. 3.7 – Exemple de représentation de la propriété de transformation de type **rotation** dans l'espace

3.6.2.3 Homothétie

L'homothétie est une forme d'évolution spatiale définie par un point et un rapport (Fig.3.8). Elle se décompose en trois phases :

- Apparition du phénomène
- Validité du phénomène sur la région de l'espace décrite par l'homothétie
- Disparition du phénomène

Cependant, cette évolution spatiale peut être précédée ou suivie d'autres transformations qui définiront son passé et son futur.

L'évolution spatiale de type translation est notée : Ho.

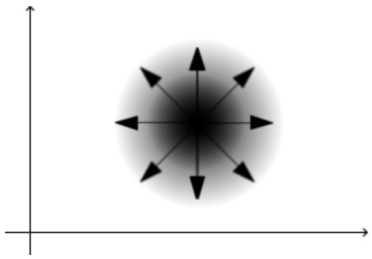


FIG. 3.8 – Exemple de représentation de la propriété de transformation de type **homothétie** dans l'espace

3.6.2.4 Propriété de transformation

Une transformation de base $T_{base} \in \{Tr, Ro, Ho\}$ entre deux événements E_1 et E_2 est notée $T_{base}(E_1, E_2)$. Si un phénomène existe alors il est spatialement et temporellement valide. Nous

pouvons compléter la propriété [3.1] :

Propriété 3.6.2

Soit $E\langle I_E, SP_E, S_E\rangle$ *un événement et* $Ph\langle O, T, S\rangle$ *un phénomène où* S_E *et* S *partagent une similarité sémantique,*

$$si\ E\langle I_E, SP_E, S_E\rangle \in Ph\langle O, T, S\rangle\ alors\ on\ a$$

$$t{-}connected(I_E, I_O) \wedge sp{-}connected(SP_E, SP_T) \tag{3.2}$$

I_O *et* SP_T *sont respectivement l'union des intervalles temporels inclus dans* O *et l'union des entités spatiales et leur transformations définies dans* T*.*

Ces différentes transformations peuvent être combinées pour engendrer de nouvelles formes d'évolution spatio temporelles. En plus des relations qualitatives classiques, nous disposons alors de moyens pour décrire les relations d'évolution entre les événements.

3.7 Conclusion

Dans ce chapitre, nous avons défini les représentations des entités temporelles et des entités spatiales de notre modèle. ceci nous a permis de formaliser la représentation des événements en définissant leurs trois composantes : temporelle, spatiale et sémantique. Nous avons également introduit la notion de relation de similarité sémantique entre deux événements. Nous avons ensuite défini les évolutions temporelles et spatiales avec les notions d'occurrence et de transformation des phénomènes.

Nous disposons maintenant des éléments qui constituent le socle de base de notre modèle temporel, spatial et sémantique. La représentation des événements est maintenant achevée. Grâce aux propriétés formelles que nous avons introduites, nous allons étudier comment utiliser les caractéristiques de ce modèle afin de proposer des possibilités de raisonnement ayant pour objectif la découverte de nouvelles relations entre les événements.

Dans le chapitre suivant nous allons étudier plusieurs points :

- la découverte d'émergence d'un phénomène
- l'agrégation d'un événement à un phénomène
- La découverte d'interactions entre les phénomènes
- Les associations sémantiques de phénomènes

Chapitre 4

Raisonnement spatial, temporel et sémantique

Dans ce chapitre nous étudions les méthodes pour découvrir de nouvelles relations entre les événements et les phénomènes. Nous allons d'abord nous intéresser à la découverte de contexte et l'émergence de phénomène. Ensuite nous définirons quelles sont les possibilités de raisonnement fournies par notre modèle.

4.1 Motivations

Nous cherchons à découvrir des relations entre événements issues d'informations ambiguës, nous avons pour cela défini un modèle formel afin de représenter temporellement, spatialement et sémantiquement ces informations et cette ambiguïté. Notre objectif est maintenant de développer des méthodes et mettre en place des règles sur notre modèle afin de découvrir de nouvelles relations. Ces éléments doivent permettre de répondre aux questions suivantes :

- Comment définir l'émergence d'un phénomène ?
- Comment associer un événement à un phénomène ?
- Comment associer les phénomènes entre eux ?

4.2 Émergence d'un phénomène

Savoir comment caractériser l'émergence d'un phénomène passe par la découverte du contexte de ce phénomène. En effet, deux à deux les phénomènes peuvent avoir des schémas d'évolutions temporelles et spatiales différents. En fonction de leurs caractéristiques spatiales et temporelles, les phénomènes qui partagent une même définition sémantique peuvent se comporter différemment.

Exemple

À la même période de l'année les cyclones au sud est des États Unis suivent une trajectoire régulière dont le point de départ se situe dans l'océan atlantique et la position finale se trouve à l'intérieur des terres (Figure 4.1) et partagent des caractéristiques similaires (Tableau 4.1).

	Katrina	Rita	Isidore
date de début	23/08/2005	18/09/2005	14/09/2002
date de fin	31/08/2005	26/09/2005	27/09/2002
durée	8 jours	8jours	13 jours
vitesse maximum du vent	278 km/s	287 km/s	204 km/s
catégorie	5	5	2

TAB. 4.1 – Caractéristiques des ouragans Katrina, Rita et Isidore [230]

Exemple

Même si on distingue des activités régulières de tremblements de terre proches des frontières entre les plaques tectoniques (Figure 4.2), les évolutions des comportements spatiaux et temporels et leurs caractéristiques restent difficiles à prévoir (Figure 4.3).

Comme nous l'avons vu, au sein d'un même domaine d'intérêt, l'émergence de phénomène est difficile à prévoir, cependant dans certains cas des caractéristiques peuvent être communes à tous les phénomènes partageant le même type. Ces éléments peuvent nous servir d'indice afin

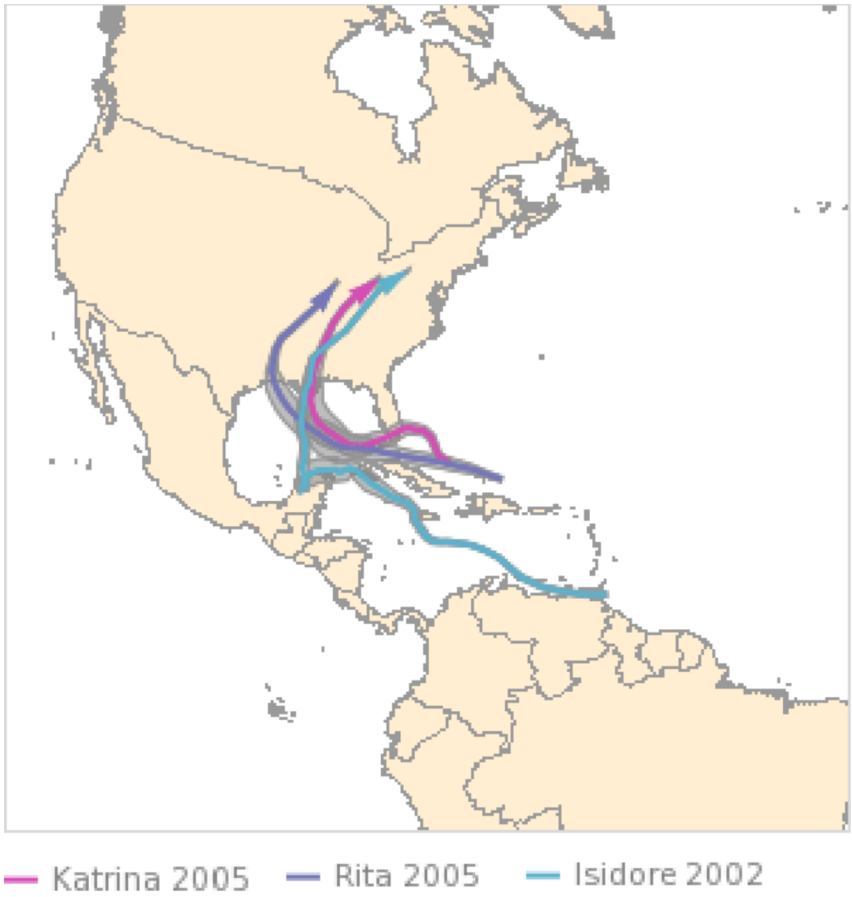


FIG. 4.1 – Comparaison des trajectoires des ouragans Katrina, Rita et Isidore [230]

de nous aider à construire automatiquement le contexte d'un type de phénomène en fonction du domaine d'intérêt.

4.2.1 Découverte automatique de contexte

Nous allons étudier quatre contextes de phénomènes dont nous avons recueillis les données depuis début 2008 afin de découvrir quels sont les comportements spatiaux et temporels intéressants à représenter (cf. le chapitre suivant pour plus de détails sur la collecte et le traitement des informations). Pour cela nous avons séparé nos données en quatre types de phénomène : cyclone, tremblement de terre, épidémie et accidents de véhicule. Nous avons alors classé les événements en les regroupant par type. Nous avons étudié la corrélation entre les événements deux à deux dans le temps et dans l'espace. C'est-à-dire que pour chaque événement appartenant à un type de phénomène donné nous avons comparé la distance temporelle et spatiale avec chaque événement appartenant au même type de phénomène. Cette interprétation permet de donner une vue d'ensemble du comportement temporel et spatial d'un type de phénomène sans avoir d'*a priori* à cause de la connaissance de sa sémantique. Nos types de phénomènes doivent être considérés comme des étiquettes, non pas comme des concepts porteurs de sens.

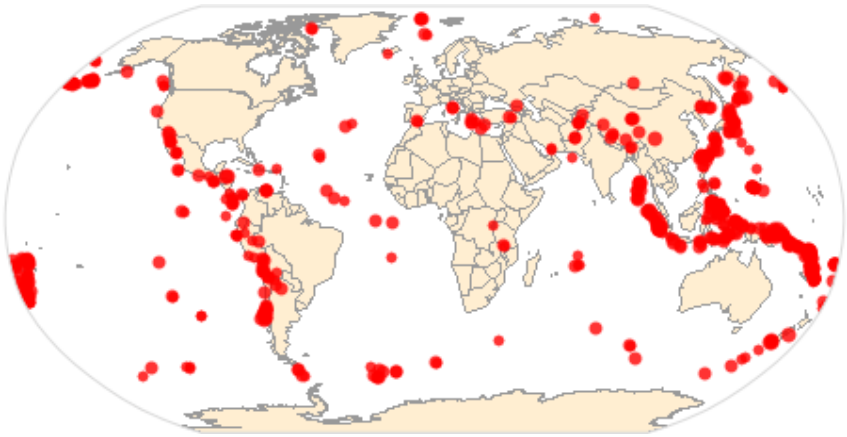


FIG. 4.2 – Répartition spatiale des tremblements de terre depuis janvier 2008 jusqu'à septembre 2010 [230]

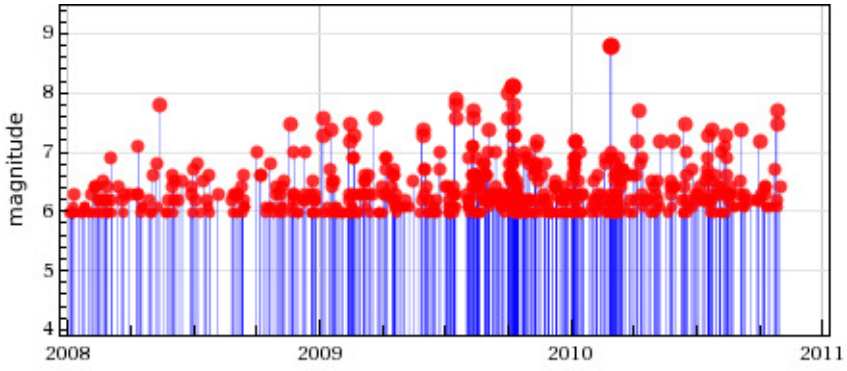


FIG. 4.3 – Magnitude des tremblements de terre depuis janvier 2008 jusqu'à septembre 2010 [230]

4.2.1.1 Étude du cas de phénomène continu dans le temps et limité dans l'espace

La disposition spatiale et temporelle des événements de la figure 4.4 rappelle fortement le caractère continu de l'évolution spatiale d'un phénomène de type *Cyclone*. Les phénomènes sont très localisés autour des régions proches de leur point d'apparition et se répètent régulièrement dans le temps. Les points situés en haut de la figure nous informent que les zones d'apparition des cyclones importants sont souvent les mêmes et qu'elles sont présentes en nombre très restreint.

FIG. 4.4 – Comparaison temporelle et spatiale deux à deux de chaque cyclone détecté sur Internet depuis début 2008 jusqu'à fin 2010

4.2.1.2 Étude du cas de phénomène à propagation temporelle et spatiale limitée

La disposition spatiale et temporelle des événements de la figure 4.5 rappelle fortement le caractère continu de l'évolution temporelle des phénomènes de type *Earthquake*. Cependant, nous considérons chaque phénomène indépendamment, nous remarquons que cette continuité est en faite une répétition due à l'échelle de la figure. Ceci correspond bien avec le comportement général des tremblements de terre ; à savoir une région spatiale limitée mais impactée à intervalle réguliers.

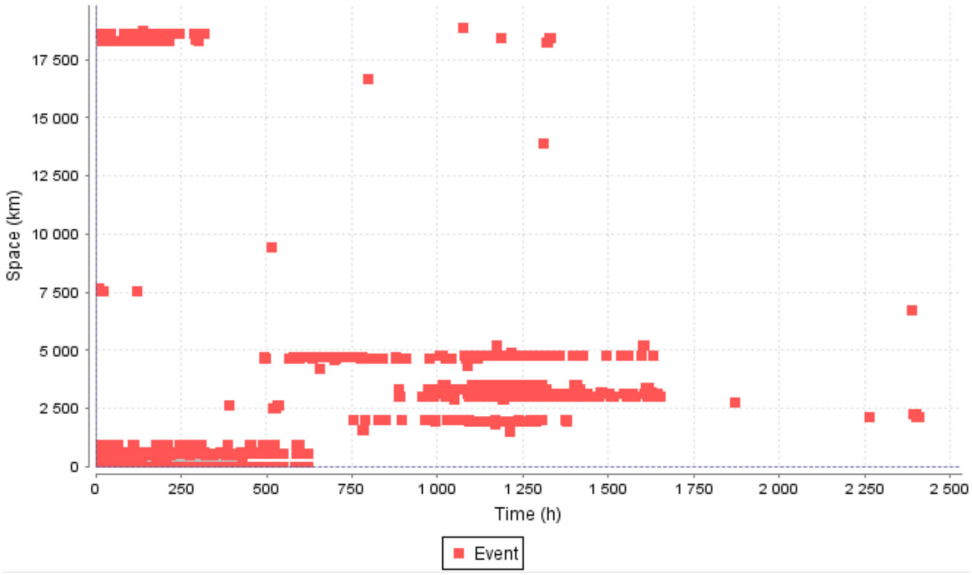


FIG. 4.5 – Comparaison temporelle et spatiale deux à deux de chaque tremblement de terre détecté sur Internet depuis début 2008 jusqu'à fin 2010

4.2.1.3 Étude du cas de phénomène à dissipation temporelle et spatiale

La disposition spatiale et temporelle des événements de la figure 4.6 rappelle le caractère aléatoire de l'apparition des phénomènes de type *Illness*. Nous remarquons que ce type de phénomène se caractérise par une brusque apparition suivie d'un développement chaotique. Au fil du temps les événements liés à ce type de phénomène vont se dissiper petit à petit pour finir par disparaître. Ce comportement correspond aux épidémies connues jusqu'ici même si certaines épidémies, par exemple en Afrique, continuent à évoluer sans présenter de tels indices de disparition.

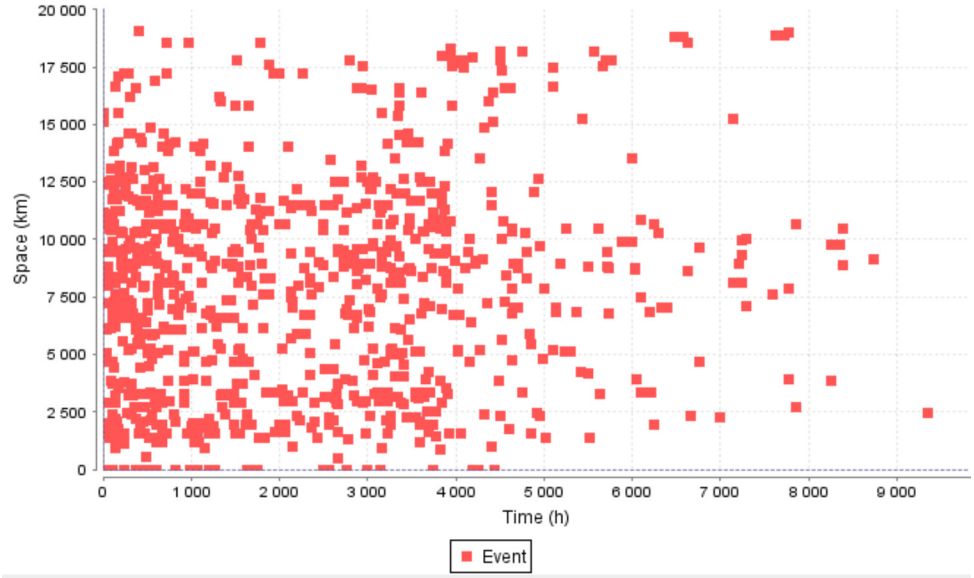


FIG. 4.6 – Comparaison temporelle et spatiale deux à deux de chaque épidémie détectée sur Internet depuis début 2008 jusqu'à fin 2010

4.2.1.4 Étude du cas de phénomène à comportement aléatoire

La disposition spatiale et temporelle des événements de la figure 4.7 peut être qualifiée d'aléatoire persistante. En effet, à tout moment, on peut voir apparaître un événement appartenant aux phénomènes de type *Vehicule accident*. Cette disposition s'explique notamment par le caractère régulier, non prévisible et instantané (à l'échelle de la figure) des accident de véhicule dans le monde.

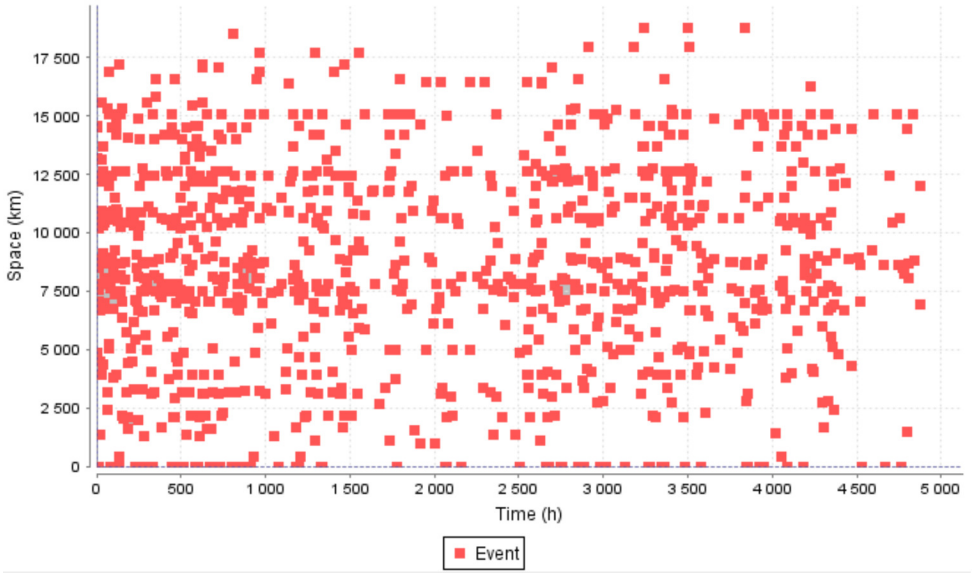


FIG. 4.7 – Comparaison temporelle et spatiale deux à deux de chaque accident de véhicule détecté sur Internet depuis début 2008 jusqu'à fin 2010

4.2.1.5 Comportement moyen

Les quatre études précédentes ne recouvrent pas l'ensemble des contextes possibles. En effet, la principale contrainte vient de la nature même des phénomènes ; ils dépendent fortement du domaine d'application. Les résultats de la corrélation de certains types de phénomènes présentent néanmoins des intérêts. Ces ensembles de phénomènes permettent de construire automatiquement un **comportement moyen** d'un type de phénomène (au sens du report d'information) sur un cadre global. Nous pouvons alors associer un phénomène émergeant avec un **comportement moyen.** Nous remarquons que certains phénomènes ne sont pas explicites à un niveau général (accidents de véhicule dans le monde), leur interprétation à une échelle plus locale (pays ou ville) apporte, en revanche, des informations dignes d'intérêt.

4.2.2 Détection de phénomènes hors normes

A partir des contextes précédents, nous pouvons définir des **comportements moyens** de phénomènes, l'intérêt est de comparer ces comportements avec les phénomènes émergeants. L'analyse spatiale et temporelle permet de définir à quel moment et quelles zones sont susceptibles d'être touchées. L'analyse du phénomène au niveau sémantique permet d'évaluer les chances d'engendrer d'autres phénomènes à partir de la situation actuelle.

Il convient de remarquer que les phénomènes que nous traitons ici sont des phénomènes extrait de l'information présente sur Internet, ils ne sont qu'une vue informationnelle du phénomène physique. Ainsi, même si les comportements temporels et spatiaux que nous avons présenté dans les précédents exemples coïncident avec une représentation naïve du comportement d'un phénomène physique, cette intuition n'est pas une règle générale et peut parfois ne pas correspondre au comportement du phénomène « réel » représenté par ce biais.

Nous disposons alors de méthodes de catégorisation des phénomènes en fonction de leurs caractéristiques. Nous allons maintenant aborder l'évolution de ces phénomène à travers l'agrégation de nouveaux événements (passés, présents ou futurs).

4.3 Agrégation d'événements

L'intérêt d'une telle modélisation est de pouvoir agréger de nouveaux événements à des phénomènes existants afin de découvrir des relations entre ces nouveaux événements et ceux connus. Pour qu'un événement appartienne à un phénomène, il existe une condition nécessaire énoncée par la propriété [2]. Si un phénomène est composé d'un seul événement alors les transformations ont une définition floue. C'est-à-dire, connaissant « l'événement de départ » ($E_d\langle I, SP, S\rangle$), nous devons considérer le domaine de définition en entier pour chaque transformation ; soit T_{base} une transformation de base d'un phénomène $Ph\langle O, T, S\rangle$ contenant uniquement E_d, alors $T = T_{base}(E_d, \star)$ où $\star$ sert de joker qui peut être remplacé par n'importe quel événement respectant T_{base} et la distribution de probabilité correspondante. Seule l'agrégation d'un futur événement E_{d+1} en relation avec E_d permettra de déterminer la direction, l'arc ou le rapport de la transformation. Ainsi, ce nouvel événement E_{d+1} remplacera $\star$.
Selon les caractéristiques du nouvel événement $E\langle I, SP, S\rangle$ reconnu comme une partie d'un phénomène $Ph\langle O, T, S\rangle$, son agrégation selon quatre règles :

Définition 4.3.1 *Règles d'agrégation*

(i) Détail de l'évolution du phénomène

(ii) Décrit l'évolution suivante du phénomène

(iii) Décrit l'évolution passée du phénomène

(iv) Nouveau départ pour le phénomène

Nous détaillons ces règles dans les paragraphes suivants. Dans les définitions suivantes SP_T désigne l'union des entités spatiales et leur transformations définies dans $T = \{Tr, Ro, Ho\}$.

4.3.1 Détail de l'évolution du phénomène

Cette règle apporte des informations concernant l'évolution courante du phénomène. Elle va détailler cette évolution en décomposant les évolutions existantes et les affiner spatialement sans les étendre dans le passé ou le futur.

Définition 4.3.2 *Règle (i)*
$\exists E_1\langle I_{E_1}, SP_{E_1}, S\rangle \in Ph$, $\exists E_2\langle I_{E_2}, SP_{E_2}, S\rangle \in Ph$ *et* $\exists T_i(E_1, E_2) \in T$ *avec* $T_i \in \{Tr, Ro, Ho\}$ *si* $I_{E_1} \leq_I I \leq_I I_{E_2}$ *et* $sp{-}connected(SP, SP_T)$ *alors* $T_i = T_j(E_1, E) \cup T_k(E_1, E)$ *où* T_j *et* T_k *sont des transformations basiques.*

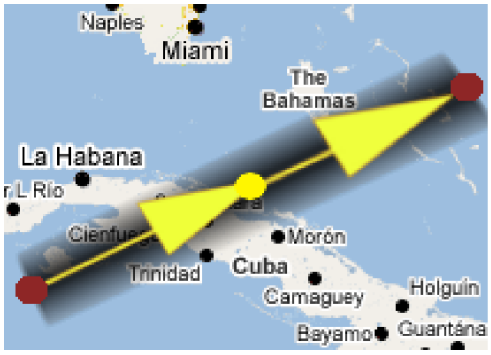


FIG. 4.8 – *(i)* le nouvel événement décrit plus précisément le phénomène[1]

4.3.2 Évolution future du phénomène

Cette règle apporte des informations concernant l'évolution future du phénomène. Elle va poursuivre ou modifier la dernière évolution connue du phénomène ajoutant des évolutions spatiales et des évolutions temporelles futures.

Définition 4.3.3 *Règle (ii)*
$\forall E_i\langle I_{E_i}, SP_{E_i}, S\rangle \in Ph, \exists E_{sup}\langle I_{E_{sup}}, SP_{E_{sup}}, S\rangle$ *tel que* $I_{E_i} \leq_I I_{E_{sup}}$, *si* $sp{-}connected(SP, SP_T)$, *on construit une nouvelle transformation* $T_n(E_{sup}, E) \in \{Tr, Ro, Ho\}$, *on a alors* $T = T \cup T_n(E_{sup}, E)$.

FIG. 4.9 – (*ii*) le nouvel événement décrit comment le phénomène va évoluer

4.3.3 Évolution passée du phénomène

Cette règle apporte des informations concernant l'évolution passée du phénomène. Elle va étendre ou remplacer la première évolution connue du phénomène en ajoutant antérieurement des évolutions spatiales et des évolutions temporelles.

Définition 4.3.4 *Règle (iii)*
$\forall E_i\langle I_{E_i}, SP_{E_i}, S\rangle \in Ph, \exists E_{inf}\langle I_{E_{inf}}, SP_{E_{inf}}, S\rangle$ *tel que* $I_{E_{inf}} \geq_I I_{E_i}$, *si* $sp{-}connected(SP, SP_T)$, *on construit une nouvelle transformation* $T_n(E, E_{inf}) \in \{Tr, Ro, Ho\}$, *on a alors* $T = T \cup T_n(E, E_{inf})$.

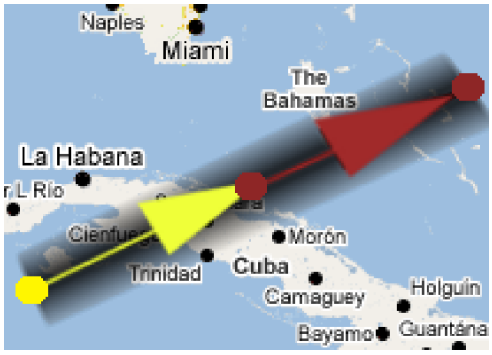


FIG. 4.10 – (*iii*) le nouvel événement décrit comment le phénomène a évolué

4.3.4 Nouveau départ d'évolution du phénomène

Cette règle apporte une nouvelle source d'information concernant l'évolution du phénomène. Elle peut décrire aussi bien une partie passée du phénomène qu'une phase qui constitue l'évolution future de celui ci par l'ajout de transformations et d'occurrences.

Définition 4.3.5 *Règle (iv)*
Dans le cas où le nouvel événement est indépendant des transformations de T, alors il représente un nouvel « événement de départ » pour le phénomène, on a $T = T \cup T(E, \star)$.

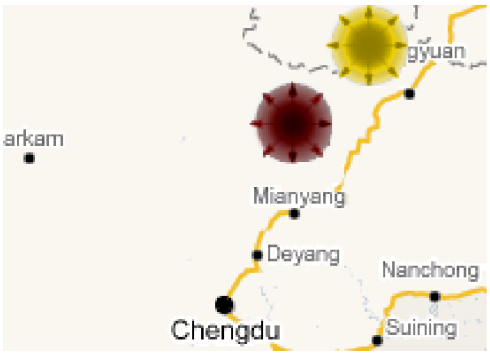


FIG. 4.11 – (*iv*) le nouvel événement décrit une autre partie du phénomène

Nous avons vu comment faire évoluer les phénomènes par l'arrivée de nouveaux événements (passés, présents ou futurs) afin d'enrichir notre connaissance de la situation. Cependant, les évolution d'un phénomène ne se résument pas à l'agrégation d'événements. En effet, comme nous allons le voir dans la partie suivante, nous devons prendre en compte les interactions entre les phénomène afin d'obtenir une représentation plus complète de la situation.

4.4 Macro Phénomène

Le concept de phénomène est utile pour décrire les évolutions des événements dont les propriétés partagent une similarité sémantique, mais il ne suffit pas pour les relations entre les événements de natures différentes. En effet, un phénomène peut contribuer à l'apparition d'un autre phénomène qui est instancié d'une propriété sémantique non similaire. Pour prendre en compte ces relations, nous introduisons le concept de *macro phénomène*. Il s'agit d'un ensemble de phénomènes en relations entre eux. Ces relations sont des liens de causalité (illustrés par la relation « *leadTo* » sur la figure 4.12) vis à vis d'un autre phénomène. L'agrégation d'un nouveau phénomène dans un macro phénomène se fait par l'agrégation d'au moins un de ses événements avec le macro phénomène en respectant les règles d'agrégation précédentes.

Définition 4.4.1 *Macro Phénomène*
Un macro phénomène MP est défini par $\forall Ph\langle O, T, S\rangle \in MP$ tel que $MP\langle \cup\{Ph\}, T'\rangle$
où T' est l'ensemble des transformations de base construites lors de l'agrégation d'un phénomène en relation avec un autre phénomène.

Exemple
Nous recevons les trois alertes en anglais suivantes, à partir desquelles nous construisons cinq événements :

1. *An earthquake in Sichuan destroyed several buildings.*
   $E_0 = E\langle I_0, Sichuan, Earthquake\rangle$,
   $E_1 = E\langle I_1, Sichuan, Destruction\rangle$
2. *More and more people are becoming sick in a village near Sichuan.*
   $E_2 = E\langle I_2, near\ Sichuan, Illness\rangle$
3. *A warehouse in Chengdu has collapsed polluting nearby rivers.*
   $E_3 = E\langle I_3, Chengdu, Pollution\rangle$,
   $E_4 = E\langle I_4, Chengdu, Destruction\rangle$

où $I_0 <_I I_1 <_I I_4 <_I I_3 <_I I_2$.

À partir de ces événements, nous recherchons des connexions temporelles et spatiales pour en déduire les phénomènes suivants :

- $Ph_0\langle\{Ra\}, \{Ho\}, Earthquake\rangle = \{E_0\}$
- $Ph_1\langle\{Co\}, \{Ho\}, Destruction\rangle = \{E_1, E_4\}$
- $Ph_2\langle\{Ra\}, \{Ho\}, Illness\rangle = \{E_2\}$
- $Ph_3\langle\{Co\}, \{Tr, Ro\}, Pollution\rangle = \{E_3\}$

Les relations de causalité, décrites dans l'ontologie de domaine (Figure 4.12 enrichie avec les relations d'occurrences et de transitions spatiales), indiquent comment construire les macro phénomènes. A partir de cet ensemble de phénomènes, nous pouvons alors associer Ph_0 et Ph_1. Ce lien est validé par les propriétés spatiales et temporelles : « *the warehouse* » se trouve dans la zone d'effet de « *the earthquake* » et sa destruction s'est produit seulement après I_0. Nous appliquons la règle *(iii)* pour agréger E_0 à Ph_1.

$$MP_0\langle\{Ph_0, Ph_1\}, \{Ho(E_0, E_1)\}\rangle$$

L'ontologie définie que les phénomène de type « *pollution* » engendre « *illness* ». Nous savons que le village où sévit la pollution est situé en aval de Chengdu et que la maladie est apparue après la pollution. Avec la règle *(ii)*, nous agrégeons E_2 à Ph_3.

$$MP_1\langle\{Ph_2, Ph_3\}, \{Tr(E_3, E_2)\}\rangle$$

Il existe une relation entre la « *factory destruction* » et « *pollution* ». Puisque leur description temporelles et spatiales sont identiques, nous regroupons MP_0 et MP_1 avec la relation entre E_4 et E_3. La règle *(i)* agrège E_3 à Ph_1 dans MP_0.

$$MP_0\langle\{Ph_0, Ph_1, Ph_2, Ph_3\}, T\rangle$$

où $T = \{Ho(E_0, E_1), Tr(E_3, E_2), Ho(E_4, E_3)\}$. Ainsi, à partir des trois alertes nous avons construit un macro phénomène grâce aux propriétés temporelles, spatiales et sémantiques. Nous avons maintenant une représentation logique des événements et de leurs relations.

4.5 Découverte d'interactions entre phénomènes

Nous avons érigé des règles d'agrégation d'événements et d'associations de phénomènes grâce aux propriétés temporelles, spatiales et sémantiques. Dans le cas idéal, une ontologie de domaine construite par un expert couvre l'ensemble des relations dont le modèle a besoin pour satisfaire les règles de découverte. Cependant, dans le cas où une ontologie des relations « *leadTo* » entre phénomènes n'est pas disponible, nous pouvons tout de même générer un ébauche d'une ontologie en supposant que l'on dispose d'assez d'informations (traduit en événements et phénomènes en accord avec le modèle) concernant un domaine d'intérêt particulier. En effet, en étudiant les

relations temporelles et spatiales entre les types de phénomènes construis selon le modèle, nous pouvons déduire des comportements émergeants qui vont, en général, dominer et par la suite définir des probabilités de relations entre les types de phénomènes.

4.5.1 Comportements dominants

Pour découvrir des comportements dominant nous avons pris notre jeu de données et nous avons regroupé les phénomènes en fonction de leurs types. Ensuite pour chaque phénomène nous avons généré un couple avec chaque phénomène qui se produit dans une zone spatiale proche du premier phénomène. Nous avons ensuite trié temporellement ces associations en écartant les associations de phénomènes présentant un trop grand écart temporel (supérieur à une journée). Nous avons alors dénombré ces associations en prenant en compte l'ordre chronologique dans lequel les phénomènes apparaissaient.

Exemple

Les associations *(Tremblement de terre*, *Inondation*) et (*Inondation*, *Tremblement de terre*) appartiennent à deux ensembles différents ; respectivement l'ensemble des relations où les « tremblements de terre engendrent des inondations » et l'ensemble des relations où les « inondations engendrent des tremblements de terre ».

Nous disposons alors d'une méthode permettant de déterminer l'existence ou non de relations privilégiées entre deux types de phénomènes. Ceci est possible car nous supposons que nous sommes en présence de types de phénomènes qui se répètent suffisamment pour pouvoir voir émerger un comportement dominant.

De plus, si nous supposons que nous connaissons les relations de hiérarchie entre les types de phénomènes, nous pouvons alors enrichir nos ensembles de relations d'un type de phénomène donné par l'intermédiaire des ensembles de relations entre les types fils de ce type de phénomène.

Exemple

L'association (*Tornade*, *Inondation*) appartient à la fois à l'ensemble des relations où les « cyclone engendrent des inondations » mais également à l'ensemble des relations où les « tempêtes engendrent des inondations ». Cette appartenance est rendue possible grâce à la relation de hiérarchie entre les types de phénomène *Tornade* et *Tempête*.

Finalement, nous obtenons le tableau 4.2 des relations entre les types de phénomènes où sont mis en valeurs les comportements dominants.

4.5.2 Probabilité d'interaction entre type de phénomènes

Nous disposons d'une méthode afin de déterminer les comportements dominants parmi les types de phénomènes de notre domaine d'intérêt. Nous avons également défini un moyen de mesurer les relations entre les phénomènes deux à deux. Nos hypothèses font que deux phénomènes indépendants apparaissant au même moment dans la même région seront associés même s'ils n'ont aucune raison de l'être. Ce biais est particulièrement visible pour le type *Vehicule-accident* (accidents de véhicules) ou *Weather* (problème météorologique) sur notre tableau. Ces types de phénomène se produisent tellement souvent un peu partout sur la planète qu'ils se retrouvent liés à la plupart des autres types de phénomènes.

Cependant, nous pouvons constater de fortes relations entre les types de phénomène *Cyclone*, *Flood, Earthquake* et *Tsunami*. Ainsi nous pouvons voir des relations émerger sans devoir effectuer une étude profonde du domaine d'intérêt. Il est nécessaire de normaliser ces résultats. En

effet, notre ensemble de données ne contient pas la même quantité d'information pour chaque phénomène. Ainsi même si une valeur peut sembler faible par rapport à la moyenne des résultats, elle reste pourtant d'intérêt pour cette association donnée.

Exemple

Par exemple la relation entre le type *Earthquake* et le *Tsunami* a une valeur de 19 qui est faible vis à vis de la moyenne du tableau mais cette valeur est la plus élevée pour l'ensemble du type *Tsunami*.

Une fois ces résultats normalisés nous pouvons alors parler de probabilité d'interactions entre type de phénomène.

4.6 Conclusion

Dans ce chapitre, nous avons étudié des méthodes de raisonnement ayant pour objectif la découverte de nouvelles relations entre les événements et entre les phénomènes. Nous disposons de méthodes de découverte semi-automatique de l'émergence de phénomène. Nous avons défini quelles sont les règles d'agrégation d'un événement à un phénomène. Nous avons également étudié l'association de phénomène au sein d'un ensemble sémantique moins restreint que nous avons appelé macro-phénomène. Finalement, nous nous sommes penché sur le problème de la découverte d'interactions entre les types de phénomènes avec une connaissance partielle d'un domaine de connaissance. Nous avons défini les critères nécessaires et suffisants afin d'appliquer ces méthodes et ces règles tout en exposant leur limites.

Dans le chapitre suivant nous allons appliquer notre modèle et ses capacités de raisonnement à la surveillance de catastrophes naturelles, nous verrons :

- Quels sont les problèmes posés par la surveillance des catastrophes naturelles sur Internet ?
- Quelle architecture adopter pour mettre au point un tel système ?
- Quels sont les résultats de notre modèle dans sa contribution au système ?

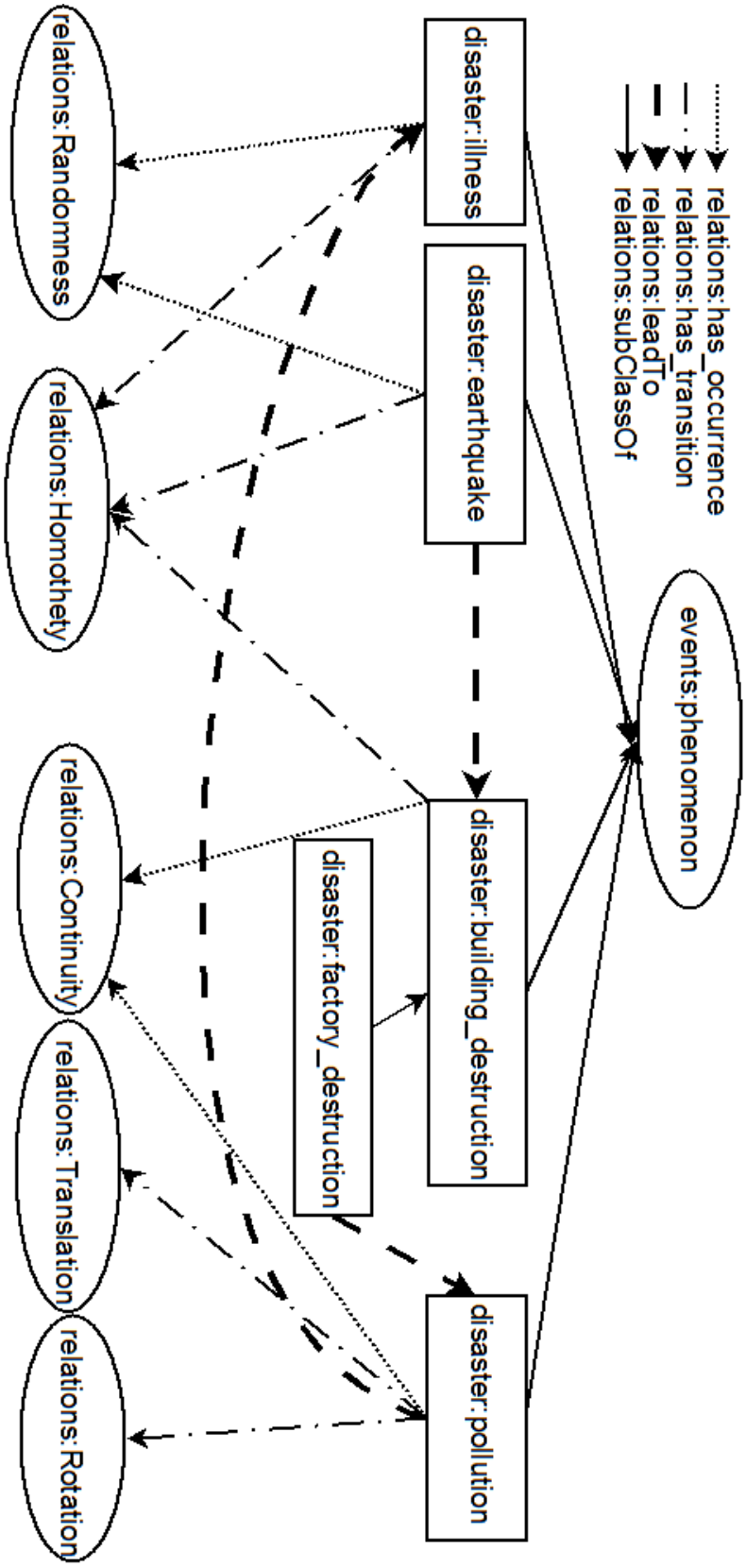


FIG. 4.12 – Partie de l'ontologie de domaine du risque.

LeadTo<-	fever	explosion	accident	bomb	volcano	earthquake	flood	cyclone	vehicle-accident	cold-wave	cholera	avalanche	tropical-depression	epidemic	monsoon	landslide	tsunami	drought	hurricane	weather	forest-fire	bird	tornado	hail	heat-wave	storm
fever	6	5	25	1	1	7	35	28	22	0	2	0	0	6	0	4	0	4	0	51	1	0	1	3	3	37
explosion	6	48	44	12	0	3	29	0	38	0	1	1	0	8	0	12	0	4	25	182	0	0	44	4	3	159
accident	0	14	55	1	0	6	14	3	48	0	1	1	0	8	0	4	0	0	4	34	5	1	5	9	1	28
bomb	8	19	11	160	0	11	11	0	10	0	0	0	0	2	0	0	0	2	0	23	0	0	7	2	0	14
volcano	1	0	1	0	10	2	15	1	1	0	0	0	0	1	0	0	0	0	0	8	4	0	4	2	2	8
earthquake	17	737	80	11	1	1190	121	88	77	0	0	0	0	39	42	41	8	1	0	266	2	19	9	10	1	197
flood	57	41	212	22	17	65	271	176	188	0	1	3	0	66	8	47	0	29	54	530	52	4	34	41	9	419
cyclone	1	24	38	0	1	30	163	2294	14	0	1	0	0	32	1	2	0	0	62	2488	0	3	1	2	6	2459
vehicle-accident	14	68	169	12	3	111	171	43	153	0	7	0	6	41	7	47	0	7	62	633	16	2	124	77	12	520
cold-wave	0	0	5	0	0	0	5	6	5	0	0	0	3	0	0	0	0	0	6	25	0	0	1	2	0	23
cholera	2	2	7	0	0	0	1	1	5	0	0	0	0	2	0	0	0	0	0	1	1	0	0	0	0	1
avalanche	1	0	4	0	0	0	0	0	4	0	0	1	0	3	0	0	0	0	0	0	0	1	0	0	0	0
tropical-depression	0	0	3	0	0	0	5	0	3	0	0	0	0	22	0	8	0	4	10	26	0	0	0	0	0	23
epidemic	17	10	43	5	0	128	55	39	39	0	3	1	0	22	0	8	0	4	12	206	1	1	20	26	3	162
monsoon	0	0	1	0	0	3	3	0	1	0	0	0	0	1	1	4	0	0	0	1	0	0	0	0	0	1
landslide	0	15	69	1	1	47	51	8	61	0	0	0	0	6	1	28	1	2	21	119	6	2	17	19	3	88
tsunami	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
drought	7	1	9	0	0	0	8	6	6	0	0	0	0	7	0	0	0	0	6	65	1	2	1	14	3	40
hurricane	0	27	36	0	0	0	25	41	36	0	0	0	0	28	0	2	0	0	266	666	3	3	28	6	6	663
weather	22	14	90	4	7	21	88	12	83	0	2	0	3	21	5	22	0	5	17	174	17	0	26	24	1	121
forest-fire	2	11	24	0	1	8	28	0	17	0	0	0	0	6	1	6	0	2	0	55	3	0	13	16	2	36
bird	4	0	3	0	0	6	10	16	2	0	0	1	0	5	0	0	0	0	16	50	0	0	0	0	1	48
tornado	0	5	66	4	0	0	60	0	61	0	0	0	0	15	0	6	0	8	1	170	27	0	60	44	8	120
hail	6	0	44	0	0	6	45	8	40	0	0	0	3	11	0	8	0	1	6	143	15	0	50	34	1	102
heat-wave	6	1	27	1	2	2	52	0	25	0	0	0	0	8	0	2	0	2	0	36	3	2	11	4	6	29
storm	17	42	204	1	1	77	163	138	181	0	2	0	3	50	3	29	0	12	311	1117	33	5	115	46	12	1054

TAB. 4.2 – Corrélations entre les types de phénomènes avec la connaissance *a priori* et les relations hiérarchiques

Chapitre 5

Application à la surveillance des catastrophes naturelles

Sommaire

Pour illustrer notre approche, nous avons développé une plate-forme de veille de catastrophes naturelles capable de traiter des flux d'information non structurés. Cette plate-forme AGATE[2] a été développée en partie dans le cadre du projet européen CITRINE. Elle est basée sur une architecture d'intégration fournie par le WebLab [89].

5.1 Contexte

La popularité des réseaux sociaux et les nouvelles formes de communication a entraîné l'apparition de nouvelles sources d'information qu'il convient d'étudier. N'importe qui est en mesure de publier et de mettre en avant les informations qui l'intéresse. Ces comportements apparaissent comme autant de moyens de suivre l'évolution d'un sujet d'intérêt : par exemple la grippe H1N1. Cependant, le traitement automatique de ces informations reste encore à améliorer pour pouvoir définir sémantiquement un sujet d'intérêt (les implications du Tsunami de Myanmar).

Nous voulons représenter des phénomènes physiques grâce aux informations que tout le monde peut trouver sur Internet. Comment modéliser ces phénomènes, leurs évolutions ainsi que leurs impacts sur l'environnement ? De telles évolutions ont été étudiées dans plusieurs domaines tels que le comportement de foule[15], la migration d'animaux[84] ou encore la gestion de flotte de véhicules[166]. Nous voulons récupérer des nouvelles issues de pages web [176] et en extraire de l'information structurée (date, zone affectée, type de phénomène) grâce au Traitement Automatique des Langues Naturelles [87, 139]. Ainsi nous pourrons construire et traiter des vues formelles et partielles du phénomène à partir des alertes telles que :

« At the beginning of the month, a rock-slide hit a shanty town in Cairo. Dozens of houses in the Manshiyet Nasser were completely destroyed by boulders and rocks. »

Ces alertes sont en anglais, elles ne contiennent pas d'informations sur leur structure (pas de metadonnée, seulement de la ponctuation). Elles décrivent des informations spatiales et temporelles qui sont parfois ambiguës.

5.2 Plate-forme WebLab

5.2.1 Origine et présentation

La plate-forme *open source* WebLab (LGPL v2.1), hébergée sur la forge OW2[231], a été développée par EADS[233], afin de faciliter l'intégration des composants et le développement d'applications en utilisant des techniques des media-mining dans le cadre de projets collaboratifs. Grâce à ces avancées réalisées, le WebLab se positionne maintenant comme la solution de technique pour l'intégration de nombreux projets (ANR / RNTL WebContent et les projets e-WokHub, SAIMSI[234], projets européens Vitalas, VIRTUOSO et Citrine[197], des études avancées pour la DGA, ou des applications industrielles cadre de programmes de grande envergure ou pour d'autres unités commerciales du groupe EADS). Dans tous ces projets, les partenaires français et européens d'EADS ont contribué à l'évolution de la plate-forme et ont collaboré afin d'améliorer la plate-forme WebLab en fonction de leurs besoins spécifiques.

La plate-forme WebLab fournit un ensemble de services pour l'exploitation des médias (texte, image, audio et vidéo) et des solutions pour les applications de veille (commerciale, stratégique, militaire, etc.). WebLab facilite l'intégration et la combinaison de composants logiciels (COTS ou *open source*) offrant des fonctionnalités telles que la collecte de site sur Internet, l'indexation

[2] Accessible à http ://eads-vdr.no-ip.org :8041/agate-viewer

de l'information et d'analyse de texte, d'analyse sémantique, d'analyse de l'image et d'analyse vidéo, la transcription de la parole au texte, la traduction et ainsi de suite.

WebLab se réfère à trois niveaux différents (cf Fig.5.1) :

- *WebLab Core* est la base technique open source (http ://www.weblab-project.org) agissant comme un environnement d'intégration pour l'interopérabilité des composants logiciels au sein d'une architecture orientée services (SOA).
- *WebLab Services* forment un ensemble cohérent de logiciels et de services élémentaires de composants d'interface homme-machine (IHM) qui peuvent facilement être intégrés avec le *WebLab Core* dans le but de construire une application dédiée. La plupart des services mit en œuvre des COTS, composants *open source* ou développés par EADS et ses partenaires.
- *WebLab Applications* résulte de l'intégration des services WebLab avec un effort minime grâce aux mécanismes de programmation *WebLab Core*. Les applications sont construites par composition de chaînes de traitement plus ou moins complexes, des services et des composants IHM pouvant être reliés entre eux grâce à des outils d'édition graphique et WYSIWYG[3].

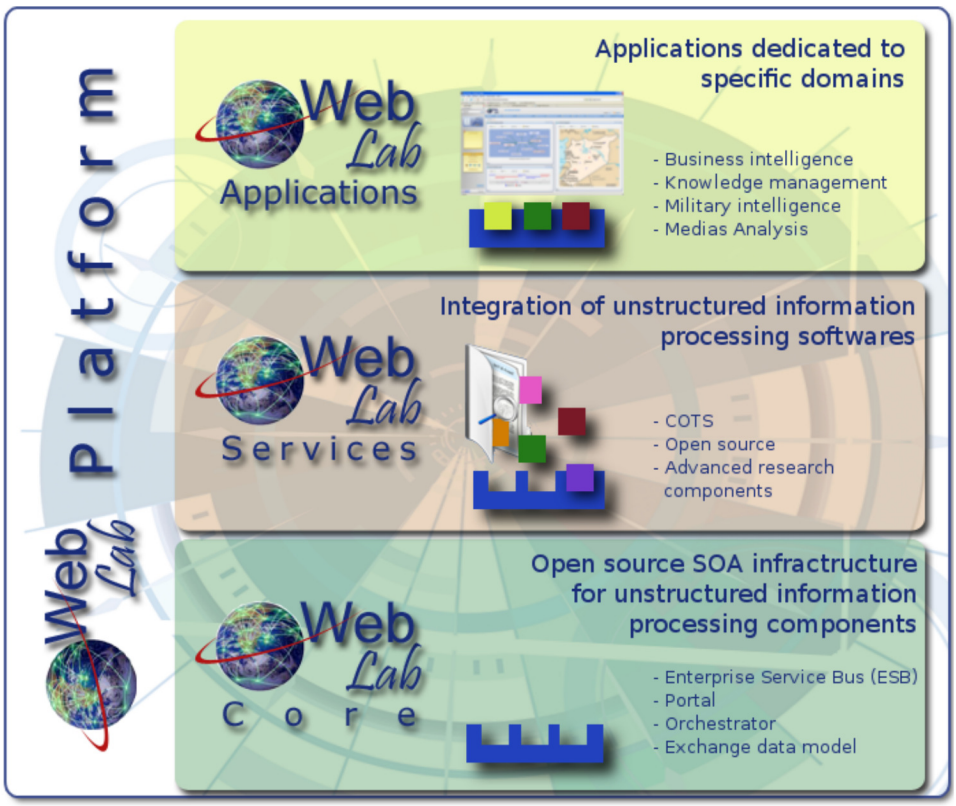


FIG. 5.1 – Les différents niveaux de la plate-forme WebLab

5.2.2 Une plate-forme d'intégration orientée service

L'architecture WebLab est organisée en plusieurs couches logiques :

[3] *What You See Is What You Get* (une traduction aproximative est : ce que vous voyez est ce que vous obtenez)

Accès : elle permet aux utilisateurs d'activer les processus d'affaires et d'interagir avec le système afin d'atteindre les processus.

Processus : elle permet de concevoir et d'exécuter l'application comme un flux de travail en utilisant un moteur d'orchestration. Elle contient un éditeur permettant de définir la chaîne de services et d'outils pour appuyer le déploiement, la configuration et la gestion des creux composants d'interface graphique.

Service d'accès : elle est composée avec des éléments de liaison qui permettent de relier les éléments multiples de logiciel.

Services aux entreprises : elle comprend tous les éléments logiciels externes qui réalisent les processus d'affaires tels que l'exploration, l'analyse et, éventuellement, des mesures de performance des composants.

Données : elle représente les sources utilisées en entrée du système et les dépôts où sont stockées les données traitées d'une exploitation ultérieure.

Infrastructure : elle contient le système d'exploitation et des outils de visualisation.

La couche d'accès permet à un utilisateur d'invoquer les services exposés par la plate-forme. Elle comprend un portail qui offre un accès unique à toutes les ressources du système (applications, services et données), contrôle les accès et gère l'identification à l'aide de session. Le portail s'occupe également de la gestion et du contrôle des éléments graphiques pour l'interface utilisateur (portlets) à travers un modèle unifié et structuré, permettant ainsi de personnaliser l'espace de travail conformément à un profil de l'utilisateur ou à un besoin de l'application. La couche d'accès est enfin complétée par un outil de conception graphique qui permet de construire des chaînes de service et de produire une description des processus en utilisant WS-BPEL standard[48].

Un bus de service prend en charge le transport des messages afin d'assurer la communication de la couche de services. Deux catégories différentes peuvent être distinguées pour les services : «technique» et «opérationnel». Chaque service est implémenté en utilisant un ou plusieurs composants logiciels et chaque composant peut implémenter des services multiples. L'intégration d'un composant logiciel externe pourrait être réalisé pour n'importe quelle couche.

5.2.3 Le modèle d'échange du WebLab

Un modèle conceptuel d'information est utilisé afin de définir un format commun d'échange (cf Figure 5.2) et de faciliter le chaînage des services de traitement : un service produit un résultat et le fournit à un service consommateur (qui va décoder les données, puis les traiter). L'orchestration est rationalisée, car les interfaces ne sont pas spécifiques aux composants mais bien aux services. Le format de données unique permettra également de réduire la complexité d'intéropérabilité entre services. Pour les mêmes raisons, la mise à jour des composants des services ne posera pas de problèmes.

Le modèle d'échange définit la grammaire commune qui, du point de vue technique sera exprimée à travers un schéma XML. Il décrit la structure et le contenu des données échangées à travers le bus de service. Afin de faciliter la communication lors de sa conception, le modèle a été formalisé avec UML. Les types XML complexes sont modélisés comme des classes d'objets qui seront utilisés dans les définitions de services en WSDL.

Les standards du Web sémantique ont été utilisés afin d'assurer la viabilité du modèle, sa compatibilité avec les composants logiciels existants impliqués dans le domaine d'application de la plate-forme WebLab, la capacité d'exploiter les ontologies de domaine existant ou de construire avec les outils d'extraction de l'information sémantique.

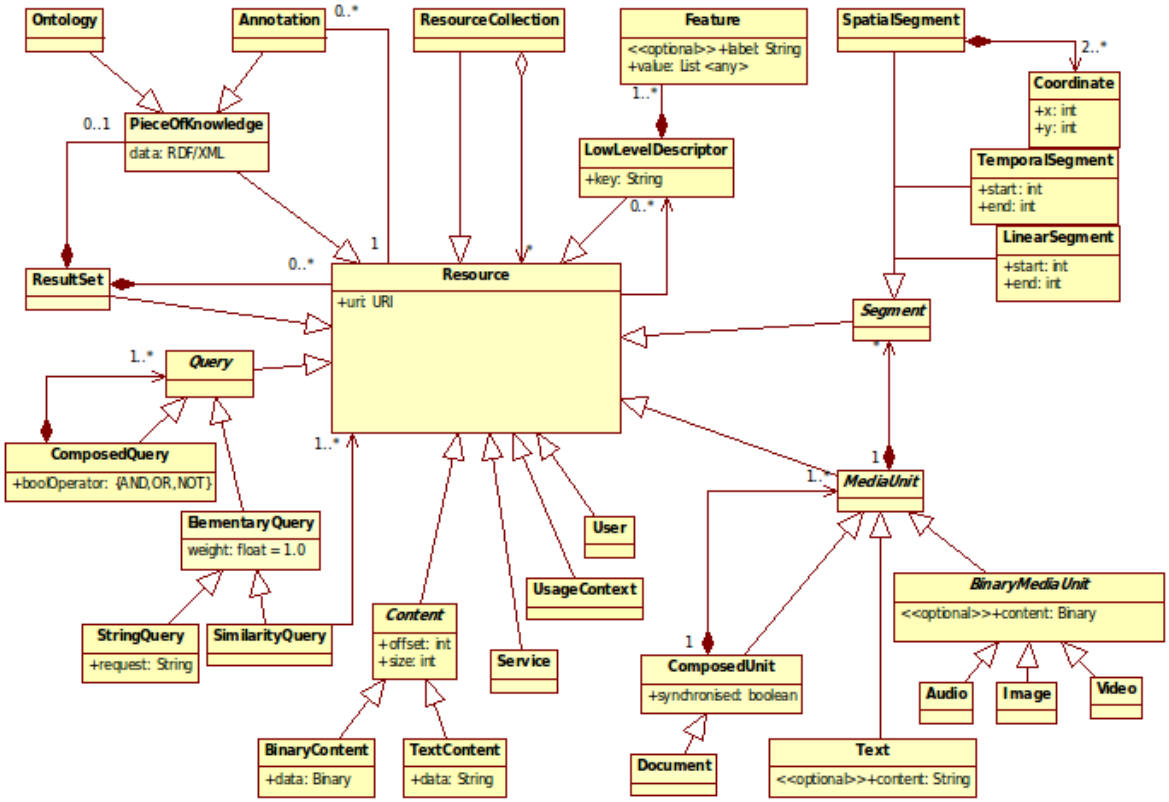


FIG. 5.2 – Modèle d'échange du WebLab

Une ressource WebLab, généralement appelée simplement une ressource, sera définie comme un objet qui a un lien avec les intérêts de l'utilisateur ou les besoins des services. Il peut être utilisé par un service sur la plate-forme et est identifié par un URI.

La notion de ressource, comprend tous les types d'entités qui doivent être traités par plusieurs services ainsi que tout type d'objet qui pourrait être utile pour un service dans une tâche spécifique : le document, le segment des documents, des fichiers de configuration des services ou des ontologies, des requêtes, etc

Sans standard commun pour décrire la structure des documents multimédia, nous avons proposé un modèle de segment multimédia documents. Il permet d'ajouter des annotations pointant vers une partie identifiée de ces documents (par exemple, les paragraphes, phrases, mots d'un document textuel).

Tous les types de ressource peuvent être annotés avec des annotations en utilisant le modèle RDF défini par le W3C. Puisque la classe « Annotation » est une ressource, elle peut également contenir des annotations. Ces déclarations peuvent être considérées comme méta-annotations décrivant, par exemple, comment et quand le premier niveau des annotations a été créé.

Le contenu des « MediaUnit » est référencé, mais pas nécessairement échangé à travers le bus afin d'éviter la surcharge et la congestion. Un protocole de communication adapté est utilisé pour transmettre le contenu entre les services. Cependant, le format commun d'échange de contenu texte reste le format XML. La classe « Segment » nous donne la possibilité de localiser des informations dans le contenu.

5.2.4 Normalisation d'interfaces

Le *WebLab Core* propose de spécifier les interfaces des services de traitements de l'information. Ces spécifications visent à normaliser et faciliter l'échange de données et l'intégration des composants. Ces interfaces sont décrites en UML puis transformé en WSDL.

Le schéma XML des interfaces génériques est inclus dans tous les services. Ainsi, la technologie Web Service permet de générer une API dans la plupart des langages de programmation (tels que Java ou C + +) qui offre des méthodes indépendantes afin de manipuler les objets du modèle d'échange.

Le modèle UML permet de décrire les interfaces abstraites et génériques pour chaque fonctionnalité d'une chaîne de traitement de l'information. Ainsi, ces interfaces peuvent être instanciées pour construire un service que nous voulons mettre en œuvre.

En suivant cette approche, neuf interfaces génériques ont été identifiés. L'instanciation de ces interfaces est utilisée pour définir tous les services. Ces interfaces sont définies dans le fichier WSDL qui est alors inclus dans tous les WSDL de chaque service de la plate-forme.

5.2.5 Orchestration

La plate-forme WebLab fournit les moyens de définir des collaborations de services WebLab atomiques afin de parvenir à un objectif d'affaires notamment à travers un processus complexe. Pour exécuter un processus métier, la plate-forme utilise un moteur d'orchestration basé sur le standard W3C : WS-BPEL. WS-BPEL décrit les interactions entre de multiples services à fournir une valeur supérieure services. L'ajout d'un niveau orchestration à la plate-forme permet un couplage lâche des services élémentaires.

WS-BPEL permet la description d'un processus métier complexe en utilisant des opérations standards telles que les appels de service (en utilisant le système de bus de routage), des blocs conditionnels, boucles, des affectations de variables, etc. Le moteur d'orchestration peut chaîner des services WebLab et permettre leurs interactions en fonction du programme WS-BPEL. Le processus résultant utilise chaque service sur demande, sans se soucier de leur mise en œuvre réelle. Cela signifie que la chaîne WS-BPEL est exposée et a été requêté à l'extérieur du bus exactement de la même manière comme une autre de service. En outre, un processus appelle un service ou une sous chaîne en utilisant le même mécanisme (il n'y a pas de différence entre le service unitaire classique et l'invocation du processus). Les chaînes simples vont exploités des services unitaires et peuvent être exposés en tant que service WebLab en utilisant les interfaces génériques. Les processus complexes, comme une chaîne d'indexation, peut utiliser des chaînes en chaînes de sous et / ou des services unitaire.

Grâce à la composition de services multi-niveaux et la capacité du bus de sélectionner dynamiquement les services à l'exécution, la plate-forme est assez souple pour s'adapter à des besoins variés. Il est possible de développer une chaîne de traitement basée sur les services existants WebLab en précisant à un éditeur graphique WS-BPEL afin de générer un programme. Le modèle WebLab est un socle commun et générique destiné à faciliter la création de la chaîne en permettant des «pipelines» de processus où une réponse du service peut être utilisée directement en tant que requête pour le service suivant.

L'ajout, la mise à jour ou la suppression d'un service n'a pas d'impact fonctionnel sur les processus des traitements dans leur ensemble. La plate-forme a juste besoin de mettre à jour un processus composite en utilisant la capacité fourni par WS-BPEL pour attribuer dynamiquement la référence de service.

5.2.6 Portail

La plate-forme WebLab intègre également un portail Web offrant à l'utilisateur un accès unique à un panel de ressources et services disponibles dans un environnement de travail personnalisable. Dans le cas du WebLab, le portail s'appuie sur la technologie portlet.

Une portlet est un composant unitaire produisant une partie de contenu HTML ou XML qui peut être affiché dans un portail. L'ensemble des parties du contenu créé par les portlets d'un portail est agrégé sous la forme d'une page Web unique. La figure 5.4 montre un exemple d'une IHM de type portail.

La mise en place d'un tel portail s'appuie sur des portlets qui offrent des interfaces utilisateur dans le portail. Chacune de ces portlets est programmée en Java en suivant les recommandations de la JSR 168 et JSR 286 [101], qui garantit la portabilité des portlets dans chaque portail respectant ces normes. Cela permet, dans le cas de la plate-forme, la planification d'un éventuel changement de portail sans avoir à reconsidérer les IHM.

La couche de "Service" de la plate-forme WebLab comprend une bibliothèque de portlets que l'on peut utiliser pour construire une interface homme-machine par la simple réunion de portlets existantes. Cette approche du développement d'IHM par la composition est comparable à l'approche utilisée pour agréger les services élémentaires dans les chaînes de traitement. Les portlets sont considérées comme des services de commande et/ou d'affichage : elles permettent un accès contrôlé aux services «opérationnels», le déclenchement de leur utilisation et l'affichage de leurs résultats. la construction d'une *WebLab Application* correspond alors à la définition d'une orchestration de services «opérationnels» et à des portlets.

5.3 Architecture de l'application

5.3.1 Chaîne de traitement

Nous avons construit une chaîne de traitement de l'information à partir des outils des domaines de Recherche d'Information et du Traitement Automatique des Langues Naturelles. Cette chaîne est composée de quatre étapes : collecte, extraction d'entités, référencement géographique et indexation. Nous nous servons du format WebLab pour représenter les documents et leurs metadonnées, par exemple : le titre, la source, les instances temporelles, les instances géographiques, les événements et leur relations.

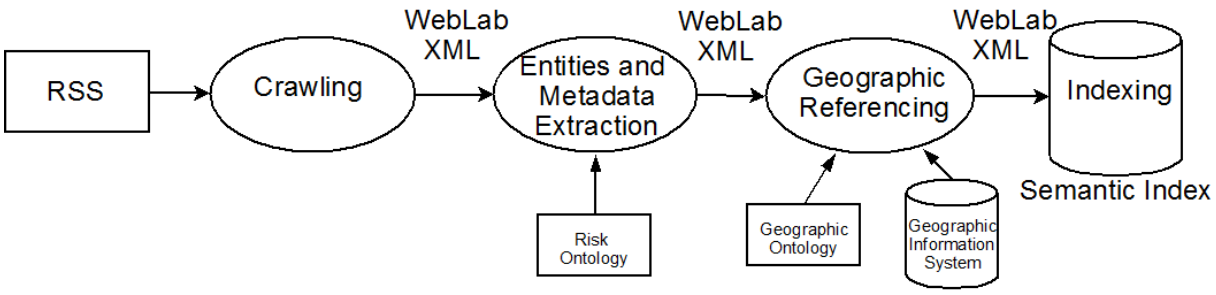


FIG. 5.3 – Chaîne de traitement Agate.

5.3.1.1 Collecte d'alertes

Nous recevons des alertes à partir des flux d'informations sur Internet mis à jour en continu. La récupération de ces informations grâce aux flux RSS nous garantit un accès rapide aux alertes récentes. Cette méthode nous permet d'accéder de manière transparente à des sources d'information hétérogènes : blogs, forums, journaux. Ces flux sont composés de metadonnées et de textes libres en anglais.

Exemple

Voici un exemple d'alerte que nous devons traiter :

« At the beginning of the month, a rock-slide hit a shanty town in Cairo. Dozens of houses in the Manshiyet Nasser were completely destroyed by boulders and rocks. »

5.3.1.2 Extraction d'entités

Les entités extraites à partir de ces textes nous permettent de construire des entités spatiales, temporelles et sémantiques. Les entités d'intérêt sont extraites des flux d'information grâce à une "pipeline GATE" [45]. Quelques entités sont extraites à partir des metadonnées selon la taxonomie Dublin Core : titre, date, source... La majeure partie des entités extraites proviennent du traitement du texte libre décrivant l'alerte. Les entités sont extraites suivant plusieurs domaines : catastrophes naturelles, victimes de catastrophes, entités spatiales et temporelles.

Exemple

Les entités extraites de l'alerte sont marqués en gras :

*« **At the beginning of the month**, a **rock-slide** hit a shanty **town in Cairo**. Dozens of houses **in the Manshiyet Nasser** were completely **destroyed** by boulders and rocks. »*

5.3.1.3 Référencement géographique

Une fois structurées, ces entités sont reliées aux instances de l'ontologie GeoNames[4]. Cette ontologie contient des propriétés de hiérarchies entre les entités géographiques (Rouen est une ville en France).

Exemple

Les entités géographiques absolues extraites de l'alerte précédente associées à des coordonnées géographiques (latitude, longitude) :

- Cairo : 30.063, 31.25
- Manshiyet Nasser : 30.032, 31.27

5.3.1.4 Indexation

Cet ensemble d'information est enfin indexé dans une base de données sémantique.

5.3.2 Interface

Cette application dispose d'une interface graphique (Fig. 5.4) destinée à la gestion et la recherche d'alertes. Elle permet donc à la fois de se tenir facilement au courant des dernières

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

catastrophes naturelles et technologiques et de rechercher parmi les informations capitalisées à des fins d'archivage.

Après traitement, les flux d'informations pertinents sont sélectionnés pour l'affichage selon le profil de l'utilisateur (les ouragans dans les îles Caraïbes en 2008 par exemple). Lorsqu'un utilisateur clique sur une alerte, sa description complète est présentée avec une mise en relief des entités d'intérêt (géographiques, catastrophes naturelles et nombre de personnes touchées). Celui ci peut alors demander quels sont les événements qui partagent des relations selon leur proximité spatiale, temporelle et sémantique.

5.3.2.1 Filtrages géographiques

En plus de la recherche textuelle standard par mots clés, les utilisateurs peuvent restreindre les alertes reportées par le système grâce à une taxonomie des catastrophes et une liste de pays d'intérêt (Fig 5.5). Les requêtes géographiques sont dessinées à l'aide de polygones sur la carte.

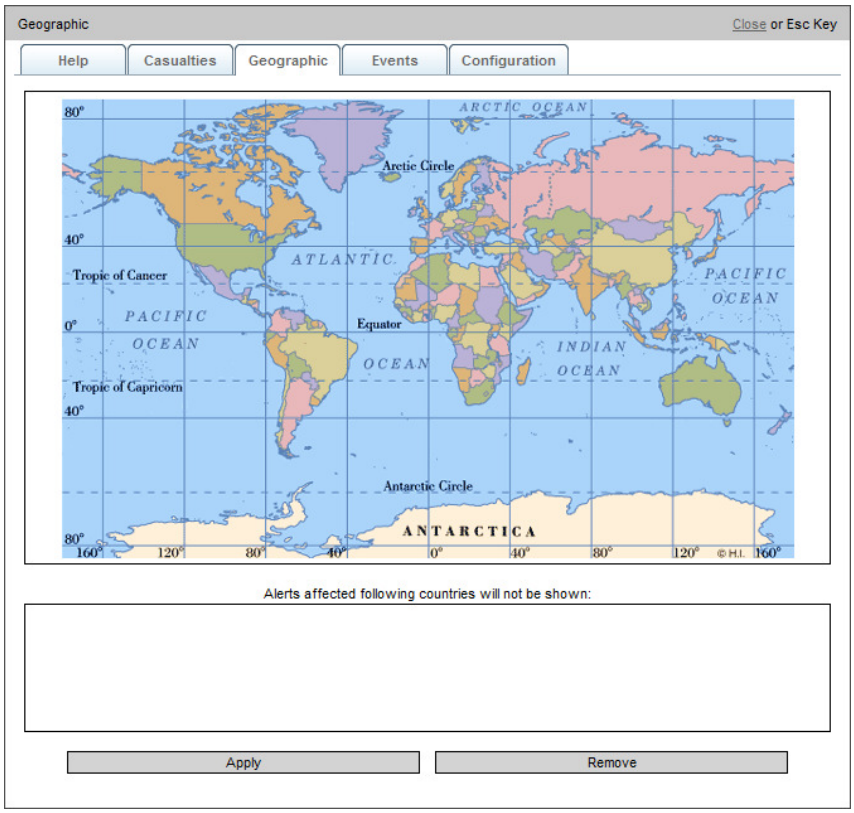


FIG. 5.5 – Filtre géographique par nation

5.3.2.2 Filtrage en fonction des victimes

Un besoin exprimé par l'Organisation Non Gouvernementale (ONG) AMI est de restreindre les alertes affichées en fonction du nombre de victimes. En effet, cette ONG est spécialisée dans l'apport d'aide médicale lors d'épidémies ou de grand problème sanitaire ou de nutrition. Afin d'éviter d'être « pollué » d'alertes non satisfaisantes (par exemple l'apparition d'un ouragan au milieu de l'océan atlantique à 1000km de la plus proche habitation) nous avons mis à disposition un système de filtre en fonction du nombre de victimes frappées par la catastrophe (Fig 5.6).

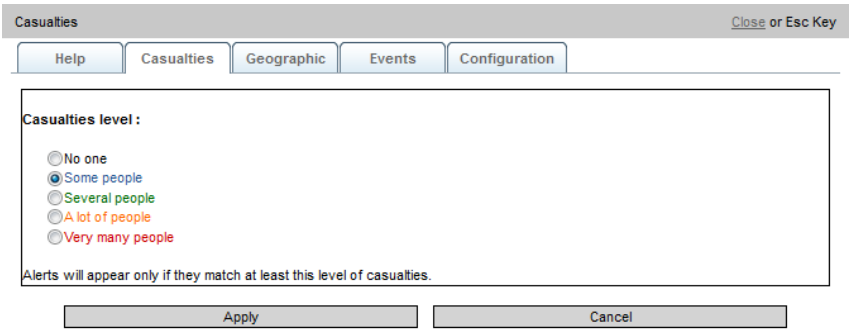


FIG. 5.6 – Filtre en fonction du nombre de victimes

5.3.2.3 Filtrage par type de phénomène

Le filtrage par type de phénomène nous a été demandé après s'être aperçu qu'un nombre important d'alerte concernant les accidents de véhicules soient remontés par le système. En effet, l'ONG spécialisée dans les traitements médicaux ne dispose pas de la structure adaptée afin de répondre à ce type de catastrophe. Nous avons donc mis au point un système de filtres par type de phénomène (Fig 5.7) qui est basé sur les différents concepts définis dans l'ontologie de domaine. Ainsi en déactivant certains types de phénomènes, ils n'apparaîtront plus dans la liste des alertes récentes. Il faut noter que les types de phénomènes fils de types de phénomènes désactivés n'apparaîtront pas non plus.

FIG. 5.7 – Filtre par type de phénomène

5.4 Expérimentations

Pour valider notre modèle, nous avons mis en œuvre la chaîne de traitement précédente afin de récupérer des alertes des trois sites d'informations Reuters, GDACS et RSOE[5]. Nous avons construit et rempli une base de données de plus de 20000 alertes depuis Avril 2008. Nous avons sélectionné, au hasard, un ensemble de 1500 alertes et pour chacune nous nous sommes assurés que l'extraction et le référencement des entités était correcte. Nous avons défini deux requêtes à tester sur cet ensemble d'alertes validées : « *Tainted milk China end of 2008* » (a) et « *Floods Myanmar May 2008* » (b). Nous comparons les résultats retournés par notre modèle implémenté sans (modèle I) et avec (modèle II) les propriétés sémantiques. Ces résultats sont présentés dans le Tableau 5.1 par liste de phénomènes triés par date d'apparition.

modèle *I*	modèle *II*
Babies killed by tainted milk in China Two deaths sentenced over tainted milk	Babies killed by tainted milk in China Arsenic contamination in Guangxi Two deaths sentenced over tainted milk

Tainted milk China end of 2008 (a)

modèle *I*	modèle *II*
Red flood alert in Myanmar Devastation on Myanmar aid mission	Tropical Storm - Asia - Myanmar Red flood alert in Myanmar Devastation on Myanmar aid mission Reports of malaria outbreaks

Floods Myanmar May 2008 (b)

TAB. 5.1 – Titres des phénomènes découverts pour chaque requête (a) et (b)

Ces résultats montrent que le modèle I retrouve des phénomènes intéressants et le modèle II retrouve au moins les mêmes. Cependant le modèle II a mis en avant des phénomènes négligés par le modèle I. Les résultats donnent plus de détails pour la requête (a) et une meilleure description des causes et conséquences pour la requête (b). Sur ces requêtes, le modèle étendu avec les propriétés sémantiques produit une meilleure description du désastre et de son contexte.

5.5 Conclusion

Dans ce chapitre, nous avons présenté quels sont les problèmes posés par la surveillance de catastrophes naturelles sur Internet. Nous avons introduit une architecture générique adaptée au besoin posé par la mise en place d'un tel système. Nous avons fourni une interface utilisateur personnalisable et facilement accessible. Les résultats des expérimentations menées sur le modèle et ses capacités de découverte de relations non triviales ont montré des résultats encourageants malgré des problèmes d'ambiguïté de l'information collectée.

[5]http ://www.reuters.com, http ://www.gdacs.org, http ://hisz.rsoe.hu

FIG. 5.4 – Interface Graphique.

Conclusion et perspectives

Bilan

Nous avons vu que l'interprétation des propriétés sémantiques en plus des propriétés temporelles, spatiales d'entités dans un texte permet de découvrir des relations insoupçonnées entre les événements décrits. Ceci est possible grâce à la représentation des phénomènes à travers une ontologie contenant des propriétés d'occurrences temporelles et de transformations spatiales. Ce modèle s'applique particulièrement bien aux phénomènes physiques tels que les catastrophes naturelles.

Nous avons abordé les problèmes de *représentation de l'information* à partir de sources ouvertes afin de mieux connaître les contraintes et les hypothèses liées aux types d'informations rencontrées. Nous avons vu que les différentes étapes du traitement de l'information que nous avons présentées (collecte, normalisation, extraction des entités, analyse des entités et indexation) ne suffissent pas toujours à caractériser de manière satisfaisante les informations. L'apport de la sémantique est une solution possible pour obtenir une représentation de l'information plus complète.

Nous avons ensuite étudié les *représentations des entités temporelles et des entités spatiales* à travers les nombreux travaux existants dans les domaines de la représentation du temps et la représentation de l'espace. Nous nous sommes surtout attardés sur la présentation des éléments en cohérence avec notre modèle afin de mettre en perspective la simplification volontaire des représentations temporelles, spatiales et sémantiques à l'intérieur du modèle pour limiter le nombre de contraintes nécessaires à son utilisation.

Nous avons alors pu commencer à donner les bases sur lesquelles se fonde le *modèle* que nous proposons. Tout d'abord nous avons défini quelles étaient les représentations du temps et de l'espace qui nous semblaient les plus adaptées à la modélisation des événements. Nous avons introduit une approche afin de faire *cohabiter des propriétés temporelles, spatiales et sémantiques*. Ainsi nous avons pu définir la notion d'événement au sein de notre modèle. En étudiant les évolutions des événements dans le but de mieux comprendre les relations qui les liaient. En nous inspirant et en étendant des travaux existant, nous avons introduit la notion de phénomène qui englobe la notion d'événement et dispose de propriétés d'*occurrence temporelle*, de *transformation spatiale* et de *similarité sémantique* propre à notre modèle.

Grâce à ce modèle nous avons décrit en détail des méthodes pour regrouper des connaissances sur les propriétés temporelles, spatiales et sémantiques d'un domaine. Ce cadre, qui est fortement dépendant du domaine, permet de faire émerger les différents types de phénomènes. Nous avons alors défini des *règles d'agrégation* d'un nouvel événement à un phénomène. L'idée est de pouvoir disposer de règles explicites pour mettre à jour facilement un phénomène lors de l'arrivée d'informations. Ces « nouvelles » informations peuvent caractériser le phénomène quelque soit la période de temps ou la région spatiale considérées. Ainsi un événement peut décrire aussi bien le passé, le présent ou le futur du phénomène auquel il est agrégé. Nous avons alors réussi à définir

une *relation temporelle, spatiale et sémantique entre événements*. Cependant, cette relation est limitée par sa propriété de similarité sémantique entre les événements.

Nous avons alors proposé une solution à ce problème en définissant *les macro phénomènes*. Cette notion va permettre de définir des relations entre des phénomènes de types différents. Ainsi, nous avons pu définir des *principes de raisonnements* à partir du modèle afin de déterminer des *relations , a priori, non explicites entre les événements*.

Finalement, nous avons présenté une application des travaux réalisés durant la thèse à travers la description d'un projet de suivi des catastrophes naturelles sur internet : *Agate*. Nous avons introduit la plate-forme d'intégration et les différents composants pour construire la chaîne de traitement des alertes et récupérer de l'information structurée, correspondant au modèle présenté, à partir de l'information non structurée récupérée dans les flux d'alertes en anglais. Cette application nous a permis d'effectuer des expérimentations sur des données réelles qui ont montré l'intérêt du *modèle pour la découverte de relations entre événements*.

Perspectives

Les travaux que nous avons entrepris au cours de la thèse ont montré des résultats encourageants mais surtout ils ont permis de mettre en avant des points difficiles qu'il est intéressant d'aborder pour confirmer la mise en $\frac{1}{2}$uvre du modèle :

- Les problèmes liées sources d'information sur Internet
- La précision de entités extraites
- La représentation du vague à plusieurs niveaux
- Les limites des méthodes de découverte semi-automatique d'information
- Les règles de découvertes de relations plus complexes

Nous allons expliciter, point par point ces éléments.

Comme nous l'avons vu précédemment, le premier point concerne les sources d'information qui posent de nombreux problèmes quand à leur disponibilité ou la crédibilité qu'il faut leur accorder. Ce sujet de recherche est complexe ; il ne semble pas y avoir de solution parfaite et chaque cas apporte son lot de problèmes spécifiques à traiter. Même s'il existe des méthodologies, des *framework* ou des bonnes pratiques à mettre en $\frac{1}{2}$uvre qui sont publiées dans la littérature, la prise en compte d'une seule information erronée, même involontairement, peut suffire à mettre en échec une règle, une condition ou une hypothèse qui entraînera potentiellement un comportement incohérent. Si la prise en compte de multiples points de vue sur une information peut amener à la remettre en question, il reste aussi simple de propager une information erronée qu'une information correcte.

Le second point traite de la précision des entités extraites. En effet, lors des expérimentations, nous avons remarqué que chaque extraction d'entité nommée se distinguait par un problème spécifiques. Nous avons validé les extractions des entités temporelles, des entités spatiales et des entités de type phénomène pour chaque alertes. Les entités temporelles posent des problèmes de contexte et de standardisation ; certaines dates sont affichées selon l'heure locale sans préciser le fuseau horaire. D'autres sont trop vagues (« *In the last month there had been three successives earthquakes* ») ou font références des éléments dont le système n'a pas connaissance (« *the day before the disaster* »). Les entités spatiales subissent également ces problèmes de contexte et de standardisation mais en plus s'ajoute un ambiguïté sur la nature même de l'entité. Les noms de villes sont parfois empruntés aux noms propres ou aux noms communs même si il existe des indices sur leur nature (par exemple ; un nom de ville commence souvent par une majuscule et il est souvent en relation avec une catégorie de verbes caractéristiques). Les entités de type phénomène

sont un cas à part, car les problèmes rencontrés lors de l'extraction sont dépendants du domaine considéré. Dans nos travaux, nous n'avons pas attaché d'importance aux relations entre les événements lors de l'extraction des entités. En plus de la mise en place d'une méthodologie à base de grammaires d'extraction d'événements, il pourrait être intéressant de considérer les retours du modèles pour prendre en compte des relations entre les entités composant les événements.

L'ambiguïté de l'information est un problème de fond qui est apparu avec le premier chapitre et n'a pas cessé de réapparaître sous diverses formes : extraction des entités nommées, représentation temporelle par des intervalles flous, représentation spatiale par des relation qualitatives ou dans une moindre mesure la propriété de similarité sémantique que nous avons utilisée. Bien que caractérisée, cette ambiguïté ne se retrouve pas dans les règles d'agrégation que nous avons définies mais il pourrait être intéressant de donner une mesure de confiance pour les événements. Nous aurions alors des ensembles d'événements valués reliées entre eux par des relations temporelles, spatiales ou sémantiques. la structure que nous obtenons ainsi est très proche d'un réseau bayésien. L'examen du modèle présenté en parallèle des propriétés sur les réseaux bayésiens pourrait permettre de définir de nouveaux types de règles ou de remettre en question des croyances concernant tels événements ou telles relations.

Nous avons présenté des méthodes semi-automatique afin de préparer les connaissances d'un domaine afin de servir de base de connaissance sémantique au modèle. Ainsi nous avons vu que ces méthodes détectaient l'émergence de nouveaux types de phénomènes. Une limite est que les données pour obtenir des résultats intéressants sont complètement dépendant du domaine d'intérêt. Ainsi, s'il a suffit des quelques milliers d'entités nommées pour obtenir les résultats que nous avons présentés dans le cadre des catastrophes naturelles, rien ne garantit que ce nombre soit suffisant pour tout autre domaine d'intérêt. Une autre limite vient des méthodes qui supposent que les phénomènes apparaissent assez fréquemment dans les données pour constituer des points remarquables. Or cette hypothèse ne tient pas dans le cas d'alerte sur des signaux faibles, à ce moment, le type de phénomène repéré sera considéré comme du « bruit de fond ». Nous avons commencé à publier certains travaux sur les passerelles entre le modèle formel sur les phénomènes assez fréquents pour être détectés par nos méthodes et les phénomènes qui constituent les signaux faibles mais il reste encore beaucoup à faire.

Le dernier point englobe plus généralement l'ensemble des remarques précédentes qu'on pourrait résumer par la question suivante ; jusqu'à quel point modéliser formellement les évolutions de la partie virtuelle d'un phénomène physique ou de société nous renseigne-t-il sur les informations portées par ce phénomène ?

Annexe A

Tableaux de corrélations entre types de phénomènes

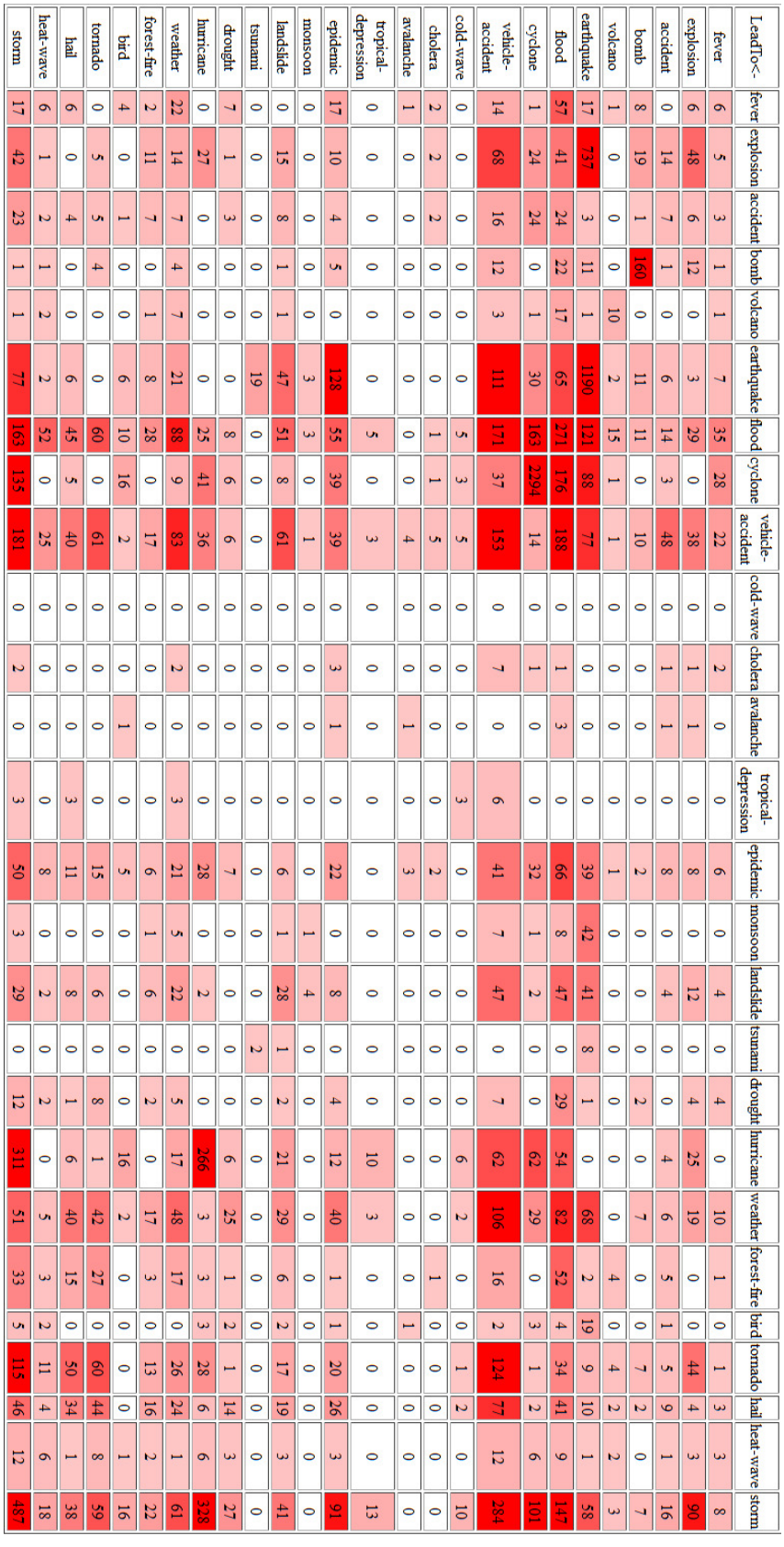
Les tableaux suivants sont les décrivent les résultats obtenus lors de la corrélation des types de phénomène entre eux (cf. Section 4.5) en ne considérant qu'une connaissance partielle du domaine d'intérêt (ici les catastrophes naturelles).

Tableau de corrélations entre types de phénomènes sans prise en compte des relations de hiérarchie

LeadTo<-	fever	explosion	accident	bomb	volcano	earthquake	flood	cyclone	vehicle-accident	cold-wave	cholera	avalanche	tropical-depression	epidemic	monsoon	landslide	tsunami	drought	hurricane	weather	forest-fire	bird	tornado	hail	heat-wave	storm
fever	6	5	25	1	1	1	10	4	22	0	2	0	0	6	0	2	0	1	0	15	1	0	1	3	3	11
explosion	6	36	32	12	0	1	5	0	26	0	1	1	0	8	0	6	0	3	1	20	0	0	7	1	3	14
accident	0	14	25	1	0	2	4	0	20	0	1	1	0	8	0	3	0	0	1	10	5	1	2	3	1	7
bomb	8	19	11	156	0	2	5	0	10	0	0	0	0	2	0	0	0	1	0	15	0	0	4	1	0	7
volcano	1	0	1	0	10	0	2	1	1	0	0	0	0	1	0	0	0	0	0	3	2	0	1	1	1	3
earthquake	2	17	18	2	1	88	14	11	16	0	0	0	0	7	1	10	3	1	0	25	2	3	4	4	1	20
flood	20	14	65	11	3	15	56	20	54	0	1	1	0	27	5	18	0	7	8	93	16	2	7	6	6	65
cyclone	1	2	14	0	1	2	12	197	12	0	1	0	0	9	1	2	0	0	16	241	0	1	1	2	1	236
vehicle-accident	14	32	114	12	3	11	44	11	98	0	7	0	2	41	4	15	0	4	5	117	14	2	24	14	11	81
cold-wave	0	0	5	0	0	0	1	2	5	0	0	0	1	0	0	0	0	0	2	10	0	0	0	1	0	8
cholera	2	2	7	0	0	0	1	1	5	0	0	0	0	2	0	0	0	0	0	1	1	0	0	0	0	1
avalanche	1	0	4	0	0	0	0	0	4	0	0	1	0	3	0	0	0	0	0	0	0	1	0	0	0	0
tropical-depression	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	2	6	0	0	0	0	0	5
epidemic	17	10	43	5	0	5	17	3	39	0	3	1	0	22	0	2	0	1	0	27	1	1	5	5	3	18
monsoon	0	0	1	0	0	1	3	0	1	0	0	0	0	0	1	4	0	0	0	1	0	0	0	0	0	1
landslide	0	4	16	1	1	10	10	0	13	0	0	0	0	1	1	16	0	0	2	19	1	0	5	4	1	10
tsunami	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
drought	5	1	6	0	0	0	3	0	5	0	0	0	0	6	0	0	0	0	0	4	1	0	0	1	1	1
hurricane	0	0	3	0	0	0	8	8	3	0	0	0	0	5	0	2	0	0	38	88	3	1	1	2	1	87
weather	4	8	39	3	1	2	18	3	35	0	2	0	1	12	2	8	0	1	4	36	8	0	3	6	1	26
forest-fire	2	11	23	0	1	4	7	0	17	0	0	0	0	6	1	5	0	2	0	12	3	0	2	3	2	6
bird	1	0	3	0	0	0	3	0	2	0	0	1	0	2	0	0	0	0	0	2	0	0	0	0	1	0
tornado	0	2	20	3	0	0	14	0	17	0	0	0	0	6	0	2	0	3	1	35	6	0	9	7	5	23
hail	1	0	17	0	0	1	8	3	14	0	0	0	1	4	0	3	0	1	2	27	4	0	6	5	1	19
heat-wave	6	1	16	1	1	0	10	0	14	0	0	0	0	8	0	0	0	2	0	16	3	1	4	4	6	9
storm	5	2	43	0	1	3	40	18	34	0	2	0	1	15	0	10	0	1	43	152	8	1	12	13	4	138

TAB. A.1 – Tableau de corrélations sans prise en compte des relations de hiérarchie

Tableau de corrélations entre types de phénomènes avec prise en compte des relations de hiérarchie limitées au premier niveau



LeadTo<-	fever	explosion	accident	bomb	volcano	earthquake	flood	cyclone	vehicle-accident	cold-wave	cholera	avalanche	tropical-depression	epidemic	monsoon	landslide	tsunami	drought	hurricane	weather	forest-fire	bird	tornado	hail	heat-wave	storm
fever	6	5	3	1	1	7	35	28	22	0	2	0	0	6	0	4	0	4	0	10	1	0	1	3	3	8
explosion	6	48	6	12	0	3	29	0	38	0	1	1	0	8	0	12	0	4	25	19	0	44	1	4	3	90
accident	0	14	7	1	0	6	14	3	48	0	1	1	0	8	0	4	0	0	4	6	5	5	5	9	1	16
bomb	8	19	1	160	0	11	11	0	10	0	0	0	0	2	0	0	0	2	0	7	0	7	2	2	0	7
volcano	1	0	0	0	10	2	15	1	1	0	0	0	0	1	0	0	0	0	0	0	4	0	4	2	2	3
earthquake	17	737	3	11	1	1190	121	88	77	0	1	3	0	39	42	41	8	1	0	68	2	19	9	10	1	58
flood	57	41	24	22	17	65	271	176	188	0	1	3	0	66	8	47	0	29	54	82	52	4	34	41	9	147
cyclone	1	24	24	0	1	30	163	2294	14	0	1	0	0	32	1	2	0	0	62	29	0	3	1	2	6	101
vehicle-accident	14	68	16	12	3	111	171	37	153	0	7	0	6	41	7	47	0	7	62	106	16	2	124	77	12	284
cold-wave	0	0	0	0	0	0	5	3	5	0	0	0	3	0	0	0	0	0	6	2	0	0	1	2	0	10
cholera	2	2	2	0	0	0	1	1	5	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0
avalanche	1	0	0	0	0	0	0	0	4	0	0	1	0	3	0	0	0	0	0	0	0	1	0	0	0	0
tropical-depression	0	0	0	0	0	0	5	0	3	0	0	0	0	0	0	0	0	0	10	3	0	0	0	0	0	13
epidemic	17	10	4	5	0	128	55	39	39	0	3	1	0	22	0	8	0	4	12	40	1	1	20	26	3	91
monsoon	0	0	0	0	0	3	3	0	1	0	0	0	0	1	1	4	0	0	0	0	0	0	0	0	0	0
landslide	0	15	8	1	1	47	51	8	61	0	0	0	0	6	1	28	1	2	21	29	6	2	17	19	3	41
tsunami	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
drought	7	1	3	0	0	0	8	6	6	0	0	0	0	7	0	0	0	0	6	25	1	2	1	14	3	27
hurricane	0	27	0	0	0	0	25	41	36	0	0	0	0	28	0	2	0	0	266	3	3	3	28	6	6	328
weather	22	14	7	4	7	21	88	9	83	0	2	0	3	21	5	22	0	5	17	48	17	0	26	24	1	61
forest-fire	2	11	7	0	1	8	28	0	17	0	0	0	0	6	1	6	0	2	0	17	3	0	13	16	2	22
bird	4	0	1	0	0	6	10	16	2	0	0	0	0	5	0	0	0	0	16	2	0	0	0	0	1	16
tornado	0	5	5	4	0	0	60	0	61	0	0	0	0	15	0	6	0	8	1	42	27	0	60	44	8	59
hail	6	0	4	0	0	6	45	5	40	0	0	0	3	11	0	8	0	1	6	40	15	0	50	34	1	38
heat-wave	6	1	2	1	2	2	52	0	25	0	0	0	0	8	0	2	0	2	0	5	3	2	11	4	6	18
storm	17	42	23	1	1	77	163	135	181	0	2	0	3	50	3	29	0	12	311	51	33	5	115	46	12	487

TAB. A.2 – Tableau de corrélations entre types de phénomènes avec prise en compte des relations de hiérarchie limitées au premier niveau

Bibliographie

[1] R. Abdulla, B. Garrison, M. Salwen, P. Driscoll, and D. Casey. The credibility of newspapers, television news, and online news. 2002.

[2] J. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11) :832–843, November 1983.

[3] J. Allen. Towards a general theory of action and time. *Artificial Intelligence*, 23 :123–154, 1984.

[4] J. Allen. Time and time again : The many ways to represent time. *International Journal of Intelligent System*, 6 :341–355, 1991.

[5] J. Allen and P. Hayes. A common-sense theory of time. pages 528–531, 1985.

[6] J. Allen and P. Hayes. A common-sense theory of time. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence, IJCAI'85*, pages 528–531. Morgan Kaufmann, 1985.

[7] J. Allen and P. Hayes. Moments and points in an interval-based temporal logic. *Computational Intelligence*, 5(4) :225–238, 1989.

[8] J. F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11) :832–843, 1983.

[9] O. Alonso, M. Gertz, and R. Baeza-Yates. On the value of temporal information in information retrieval. 2007.

[10] I. Arikan, S. Bedathur, and K. Berberich. Time will tell : Leveraging temporal expressions in ir. 2009.

[11] A. Arnold and Guessarian. *Mathématiques pour liinformatique*. Masson, masson edition, 1992.

[12] F. Bacchus, J. Tennenberg, and J. Koomen. A non-reified temporal logic. *Artificial Intelligence*, 52 :87–108, 1991.

[13] R. Baeza-Yates, C. Castillo, M. Marin, and A. Rodriguez. Crawling a country : better strategies than breadth-first for web page ordering. pages 864–872, 2005.

[14] P. Balbiani and P. Muller. Le raisonnement spatial. *Ecole thématique" Documents & Evolution" du GDR III, Le temps, l'espace et l'évolutif*, pages 33–53.

[15] M. Batty, J. Desyllas, and E. Duxbury. The discrete dynamics of small-scale spatial events : agent-based models of mobility in carnivals and street parades. *IJGIS*, (7), 2003.

[16] R. Bera and C. Claramunt. Topology-based proximities in spatial systems. *Journal of Geographical Systems*, 5(4) :353–379, 2003.

[17] T. Berners-Lee and D. Connolly. Hypertext markup language. *Internet Working Draft*, 13, 1993.

[18] T. Berners-Lee and D. Connolly. Notation 3 (n3) a readable rdf syntax. *W3C Submission, Jan*, 2008.

[19] T. Berners-Lee and J. Hendler. Scientific publishing on the semantic web. *Nature*, 410 :1023–1024, 2001.

[20] H. Bestougeff and G. Ligozat. *Outils logiques pour le traitement du temps*. Masson, Paris, 1989.

[21] F. Bilhaut and A. Widlöcher. Linguastream : An integrated environment for computational linguistics experimentation. In *Proceedings of the 11th Conference of the European Chapter of the Association of Computational Linguistics (EACL)(Companion Volume), Trento, Italie*, pages 95–98, 2006.

[22] M. Bouzid. Système de maintien de vérité dans une logique temporelle numérique et symbolique. In *Proceedings of II^{iemes} Rencontres Nationales des Jeunes Chercheurs*, page 333. AFIA, 1994.

[23] M. Bouzid, F. Charpillet, P. Marquis, and J.-P. Haton. Assumption-based truth maintenance in presence of temporal assertions. In *Proceedings of the 6th IEEE International Conference on Tools with Artificial Intelligence*, pages 492–498. IEEE Society Press, 1994.

[24] M. Bouzid and P. Ladkin. Rules for simple temporal reasoning. In *Proceedings of TIME-95, International Workshop on Temporal Representaion and Reasoning*, pages 73–88, 1995.

[25] M. Bouzid and P. Ladkin. Simple reasoning with time-dependent propositions. *International Interest Group in Pure and Applied Logic (IGPL)*, 1995. to appear.

[26] M. Bouzid and A. Ligeza. Temporal logic based on characteristic functions. In C. R. I. Wachsmuth and W. Brauer, editors, *Advances in Atrificial Intelligence, 19th Annual German Conference on Artificial Intelligence, volume 981 of Lecture Notes in Artificial Intelligence*, pages 221–232. Speinger Verlag, 1995.

[27] M. Bouzid and A. Ligeza. A temporal representation based on characteristic functions. In *Proceedings of the 8th Florida Artificial Intelligence Research Symposium, FLAIRS'95*, pages 167–172, 1995.

[28] R. Béra. *L'adjacence relative : une étude contextuelle de l'influence de l'environnement spatial dans l'appréhension de la notion de proximité*. PhD thesis, Thèse de doctorat de l'université de Rennes 1, École Navale, 2004, 2004.

[29] R. Béra and C. Claramunt. Relative adjacencies in spatial pseudo-partitions. *Conference on Spatial Information Theory (COSIT'03)*, pages 218–234, 2003.

[30] T. Bray, J. Paoli, C. Sperberg-McQueen, E. Maler, and F. Yergeau. Extensible markup language (xml) 1.0. *W3C recommendation*, 6, 2000.

[31] B. Bruce. A model for temporal references and its application in question answering program. *Artificial Intelligence*, 4 :1–25, 1972.

[32] J. Budzik, K. Hammond, and L. Birnbaum. Information access in context. *Knowledge-Based Systems*, 14(1-2) :37–53, 2001.

[33] R. Cahill. Process physics. *Process Studies Supplement*, 5 :1–131, 2003.

[34] C. Chang and H. Keisler. *Model Theory*. North Holland, 1973.

[35] E. Charniak, C. Riesbeck, and D. V. M. J. Meehan. *Artificial Intelligence Programming*. 2nd ed., Lawrence Erlbaum Associates, Hillsdale, N.J., 1987.

[36] H. Chaudet. Steel : A spatio-temporal extended event language for tracking epidemic spread from outbreak reports. In *Proceedings of KR-MED*, 2004.

[37] M. Cobb, F. Petry, and K. Shaw. Fuzzy spatial relationship refinements based on minimum bounding rectangle variations. *Fuzzy sets and systems*, 113(1) :111–120, 2000.

[38] A. Cohen, C. Bjornsson, S. Temple, G. Banker, and B. Roysam. Automatic summarization of changes in biological image sequences using algorithmic information theory. *IEEE transactions on pattern analysis and machine intelligence*, pages 1386–1403, 2008.

[39] A. Cohn and S. Hazarika. Qualitative spatial representation and reasoning : An overview. *FI*, 46(1), 2001.

[40] A. Cohn and S. Hazarika. Spatio-temporal continuity in geographic space. 2001.

[41] A. Cohn and J. Renz. Qualitative spatial representation and reasoning. *Foundations of Artificial Intelligence*, 3 :551–596, 2008.

[42] G. Corso, A. Gulli, and F. Romani. Ranking a stream of news. *WWW*, 5 :97–106.

[43] H. Couclelis. Worlds of information : The geographic metaphor in the visualization of complex information. *Cartography and Geographic Information Science*, 25(4) :209–220, 1998.

[44] S. Cox, P. Daisy, R. Lake, C. Portele, and A. Whiteside. Opengis geography markup language (gml 3.1), implementation specification version 3.1. 0, recommendation paper, ogc doc. 2004.

[45] H. Cunningham. Gate : A framework and graphical development environment for robust nlp tools and applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.

[46] H. Cunningham and al. Jape–a java annotation patterns engine. *Advances in Text Processing, TIPSTER Program Phase II*, pages 185–189, 2000.

[47] H. Cunningham, R. Gaizauskas, and Y. Wilks. A General Architecture for Text Engineering (GATE) – a new approach to Language Engineering R&D. Technical Report CS – 95 – 21, Department of Computer Science, University of Sheffield, 1995. `http ://xxx.lanl.gov/-abs/cs.CL/9601009`.

[48] F. Curbera, F. Leymann, T. Storey, D. Ferguson, and S. Weerawarana. Web services platform architecture : Soap, wsdl, ws-policy, ws-addressing, ws-bpel, ws-reliable messaging and more. 2005.

[49] A. Daude, J. Provitolo, E. Dubos-Paillard, J. Gaillard, E. Eliot, P. Langlois, and E. Propeck. Spatial risks and complex systems : methodological perspectives. 2007.

[50] M. Daum, R. Jüllig, and P. Ladkin. Approaches to planning in the Project Management Assistant. In *Proceedings of the 3rd Annual Knowledge-Based Software Engineering Conference*, pages 31–52. RADC (COES), Griffis AFB, NY 13441, 1988.

[51] A. Dauphiné. *Risques et catastrophes : observer, spatialiser, comprendre, gérer*. A. Colin, 2001.

[52] B. Davey and H. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, 1990.

[53] T. De Groeve, L. Vernaccini, and A. Annunziato. Global disaster alert and coordination system. *In Proceedings of the 3rd International ISCRAM Conference*, Newark :1–10, 2006.

[54] J. de Kleer. An assumption-based thruth maintenance system. *Artificial Intelligence*, 28 :127–162, 1986.

[55] J. de Kleer. Extending the atms. *Artificial Intelligence*, 28 :163–196, 1986.

[56] J. de Kleer. Problem soving with the atms. *Artificial Intelligence*, 28 :197–224, 1986.

[57] J. de Kleer. A general labeling algorithm for atms. In *Proceedings of the 7th National Conference on AI, AAAI'88*, pages 188–192, 1988.

[58] J. de Kleer, K. Forbus, and D. Allester, 1989. Truth Maintenace System, Tutorial SA5, IJCAI'89.

[59] M. de la Asuncion, L. Castillo, J. Fdez-Olivares, O. Garcıa-Perez, A. Gonzalez, and F. Palao. Siadex : A real-world planning approach to forest fire fighting. 2004.

[60] T. Dean. Using temporal hierarchies to efficiently maintain temporal databases. *Journal of the ACM*, 36(4) :687–718, 1989.

[61] T. Dean and D. McDermott. Temporal data base management. *Artificial Intelligence*, 32(1) :1–55, April 1987.

[62] R. Dechter, I. Meiri, and J. Pearl. Temporal constraint networks. *Artificial Intelligence*, 49 :61–95, May 1991.

[63] P. Denning, J. Horning, D. Parnas, and L. Weinstein. Wikipedia risks. *Communications of the ACM*, 48(12) :152, 2005.

[64] M. Donnelly and B. Smith. Layers : A new approach to locating objects in space. *Lecture notes in computer science*, pages 46–60.

[65] D. Dowty. *Word Meaning and Montague Grammar.* Reidel, 1979.

[66] J. Doyle. A thruth maintenance system. *Artificial Intelligence*, 12 :231–272, 1979.

[67] J. Doyle. The ins and outs of reason maintenance. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence, IJCAI'83*, pages 349–351. Morgan Kaufmann, 1983.

[68] O. Dressler. An extended basic atms. In *Proceedings of the 2th International Workshop on Non-Monotonic Reasonning, Lecture Notes in Artificial Intelligence, Springer-Verlag*, pages 143–163, 1988.

[69] O. Dressler. Extending the basic atms. In *Proceedings of the 8th European Conference on AI, ECAI'88*, pages 1535–541, 1988.

[70] O. Dressler. Problem solving with the nm- atms. In *Proceedings of the 9th European Conference on AI, ECAI'90*, pages 253–258, 1990.

[71] D. Dubois and H. P. J. Lang. Gestion d'hypothèses en logique possibiliste, un exemple d'application au dignostic. In *Actes du 10èmes Avignon, Conf. Outils, Techniques et Applications*, pages 299–313. EC2, 1990.

[72] F. Dumoncel. Un modèle d'adjacence conceptuelle pour la recherche d information géographique. *RTE 2005*, 2006.

[73] E. Grégoire. *Logiques non monotones et intelligence artificielle.* Mario Borillo et Frédéric Nef, Hermès edition, 1990.

[74] M. Egenhofer. Deriving the composition of binary topological relations. *Journal of Visual Languages and Computing*, 5(2) :133–149, 1994.

[75] M. Egenhofer, E. Clementini, and P. Di Felice. Topological relations between regions with holes. *International Journal of Geographical Information Systems*, 8(2) :129–142, 1994.

[76] M. Egenhofer and R. Franzosa. Point-set topological spatial relations. *International Journal of Geographical Information Science*, 5(2) :161–174, 1991.

[77] M. Ester, M. Groß, and H. Kriegel. Focused web crawling : A generic framework for specifying the user interest and for adaptive crawling strategies. 2001.

[78] R. Feldman and J. Sanger. The text mining handbook : advanced approaches in analyzing unstructured data. *Computational Linguistics*, 34(1).

[79] A. Flanagin and M. Metzger. Perceptions of internet information credibility. *Journalism and Mass Communication Quarterly*, 77(3) :515–540, 2000.

[80] L. Floridi. Trends in the philosophy of information. *Philosophy of information*, page 113, 2008.

[81] K. Forbus and J. de Kleer. *Building Problem Solvers*. MIT Press, 1993.

[82] A. Galton. Towards a qualitative theory of movement. 1995.

[83] A. Galton and M. Worboys. Processes and events in dynamic geo-networks. *Lecture Notes in Computer Science*, 3799 :45, 2005.

[84] D. Ganskopp. Manipulating cattle distribution with salt and water in large arid-land pastures : a gps/gis assessment. *Applied Animal Behaviour Science*, 73(4) :251–262, 2001.

[85] A. Gerevini and L. Schubert. Efficient temporal reasoning through timegraphs. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence, IJCAI'93*, pages 648–654. Morgan Kaufmann, 1993.

[86] A. Gerevini and L. Schubert. An efficient method for maintaining disjunctions in qualitative temporal reasoning. In J. Doyle, E. Sandewall, and P. Torasso, editors, *Proceedings of the Fourth International Conference, KR'94*, pages 214–225. Morgan Kaufmann, 1994.

[87] F. Gey, R. Larson, M. Sanderson, H. Joho, P. Clough, and V. Petras. Geoclef : The clef 2005 cross-language geographic information retrieval track overview. *LNCS*, 2006.

[88] M. Ghallab and M. Allaoui. Managing efficiently temporal relations through indexing spanning trees. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence, IJCAI'89*, pages 1297–1303. Morgan Kaufmann, 1989.

[89] P. Giroux and al. Weblab : An integration infrastructure to ease the development of multimedia processing applications. *ICSSEA*, 2008.

[90] M. Golumbic and R. Shamir. Complexity and algorithms for reasoning about time : A graph-theoretic approach. *Journal of the ACM*, 40(5) :1108–1133, November 1993.

[91] G. Grätzer. *General Lattice Theory*. Academic Press, 1978.

[92] P. Grenon and B. Smith. Snap and span : Towards dynamic spatial ontology. *Spatial Cognition & Computation*, 4(1) :69–104, 2004.

[93] P. Gross, S. Gupta, G. Kaiser, G. Kc, and J. Parekh. An active events model for systems monitoring. 2001.

[94] J. Gudmundsson, P. Laube, and T. Wolle. Movement patterns in spatio-temporal data. *Encyclopedia of GIS*, pages 726–732.

[95] J. Halpern and Y. Shoham. A propositional modal logic of time intervals. *Journal of the ACM*, 38(4) :935–962, October 1991.

[96] C. Hamblin. Instants and intervals. *Studium Generale*, 27 :127–134, 1971.

[97] J.-P. Haton, N. Bouzid, F. Charpillet, B. l. M-C Haton, H. Lâasri, P. Marquis, T. Mondot, and A. Napoli. *Le Raisonnement en Intelligence Artificielle*. InterEdition, 1991.

[98] B. Haugh. Non-standard semantics for the method of arguments. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence, IJCAI'87*, pages 449–454. Morgan Kaufmann, 1987.

[99] P. Hayes. The second naive physics manifesto. page 63, 1989.

[100] J. Hellendoorn. *Reasoning with Fuzzy Logic*. The Netherlands, Delft edition, 1990.

[101] S. Hepper et al. Jsr 286, portlet specification 2.0. 2005.

[102] K. Hornsby and M. Egenhofer. Modeling moving objects over multiple granularities. *Annals of Mathematics and Artificial Intelligence*, 36(1) :177–194, 2002.

[103] Z. Huang. Fuzzy temporal interval relationship based on interval-valued fuzzy sets. In *Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery*, pages 169–172, 2007.

[104] C. Hudelot, J. Atif, and I. Bloch. Ontologie de relations spatiales floues pour le raisonnement spatial dans les images.

[105] M. Huff, S. Schwan, and B. Garsoffky. The spatial representation of dynamic scenes–an integrative approach. *LECTURE NOTES IN COMPUTER SCIENCE*, 4387 :140, 2007.

[106] I. Humberstone. Interval semantics for tense logic : Some remarks. *Journal of Philosophical Logic*, 8 :171–196, 1979.

[107] G. L. J.-F. Condotta and M. Saade. Eligibilité de contraintes pour la résolution de réseaux de contraintes qualitatives temporelles et spatiales. *Atelier RTE 2007, plateforme AFIA*, 2 juillet 2007.

[108] A. Jain. Fundamentals of digital image processing. 1989.

[109] A. Java, T. Finin, and S. Nirenburg. Semnews : A semantic news framework.

[110] K. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 60(5) :493–502, 2004.

[111] C. Joubel and O. Raiman. How time changes assumptions. In *Proceedings of 9th European Conference on Artificial Intelligence*, pages 378–383, 1990.

[112] G. G. K. Kahn and. Mechanizing temporal knowledge. *Artificial Intelligence*, 9 :87–108, 1977.

[113] Y. KALFOGLOU and M. SCHORLEMMER. Ontology mapping : the state of the art. *The Knowledge Engineering Review*, 18(01) :1–31, 2003.

[114] H. Kautz and P. Ladkin. Integrating metric and qualitative temporal reasoning. In *Proceedings of the 9th National Conference on AI, AAAI-91*, pages 241–246. AAAI Press, 1991.

[115] D. Kaye. Sources of information, formal and informal. *Library Management*, 16(5) :16–19, 1995.

[116] G. Keller. The application of reason maintenace system. In *Advanced Topics in AI, Summer School*, pages 208–237, Prague, 1992.

[117] R. Kosala and H. Blockeel. Web mining research : A survey. *ACM SIGKDD Explorations Newsletter*, 2(1) :1–15, 2000.

[118] M. Koubarakis. Complexity results for first-order theories of temporal constraints. In J. Doyle and E. Sandewall, editors, *Proceedings of the Fourth International Conference, KR'94*. Morgan Kaufmann, May 1994.

[119] R. Kowalski and M. Sergot. A logic-based calculus of events. *New Generation Computing*, 4(1) :67–95, 1986.

[120] P. Ladkin. Primitives and units for time specification. In *Proceedings of the 5th National Conference on AI, AAAI'86*, pages 354–359. Morgan Kaufmann, 1986.

[121] P. Ladkin. Primitives and units for time specification. pages 354–359, 1986.

[122] P. Ladkin. Time representation : A taxonomy of interval relations. In *Proceedings of the 5th National Conference on AI, AAAI'86*, pages 360–366. Morgan Kaufmann, 1986.

[123] P. Ladkin. Time representation : A taxonomy of interval relations. pages 354–359, 1986.

[124] P. Ladkin. The completeness of a natural system for reasoning with time intervals. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence, IJCAI'87*, pages 462–467. Morgan Kaufmann, 1987.

[125] P. Ladkin. *The Logic of Time Representation.* PhD thesis, University of California at Berkeley, 1987.

[126] P. Ladkin. Models of axioms for time intervals. In *Proceedings of the 6th National Conference on AI, AAAI'87*, pages 234–239. Morgan Kaufmann, 1987.

[127] P. Ladkin. Satisfying first-order constraints about time intervals. In *Proceedings of the 7th National Conference on AI, AAAI'88*, pages 512–517. Morgan Kaufmann, 1988.

[128] P. Ladkin. Satisfying first-order constraints about time intervals. pages 512–517, 1988.

[129] P. Ladkin and R. Maddux. The algebra of convex intervals. Technical Report KES-U-87-2, Krestel Institute, 1987.

[130] P. Ladkin and R. Maddux. On binary constraint networks. *Journal of the ACM*, 41(3) :435–469, May 1994.

[131] P. Ladkin and R. Maddux. On binary constraint problems. *Journal of the ACM (JACM)*, 41(3) :435–469, 1994.

[132] P. Ladkin and A. Reinefeld. Effective solution of qualitative interval constraint networks. *Artificial Intelligence*, 57(1) :105–124, September 1992.

[133] P. Ladkin and A. Reinefeld. A symbolic approach to interval constraint problems. In J. Calmet and J. Campbell, editors, *Artificial Intelligence and Symbolic Mathematical Computing*, number 737 in Lecture Notes in Computer Science, pages 65–84. Springer-Verlag, 1993.

[134] O. Lassila and R. Swick. Resource description framework (rdf) model and syntax. *World Wide Web Consortium, http ://www. w3. org/TR/WD-rdf-syntax.*

[135] P. Laube, M. van Kreveld, and S. Imfeld. Finding remo–detecting relative motion patterns in geospatial lifelines. pages 201–214, 2004.

[136] A. Lautenschütz, C. Davies, M. Raubal, A. Schwering, and E. Pederson. The influence of scale, context and spatial preposition in linguistic topology. *LECTURE NOTES IN COMPUTER SCIENCE*, 4387 :439, 2007.

[137] B. Le Saux and N. Boujemaa. Unsupervised robust clustering for image database categorization. *Pattern Recognition*, 1 :10259, 2002.

[138] J. Lesbegueries, M. Gaio, and P. Loustau. Geographical information access for non-structured data. In *Proceedings of the ACM symposium on Applied computing*, 2006.

[139] J. Lesbegueries and P. Loustau. Structuration d'information spatiale qualitative pour la recherche d'information. 2006.

[140] J. Leveling and S. Hartrumpf. On metonymy recognition for geographic ir. In *SIGIR Workshop on Geographical Information Retrieval, Seattle*, 2006.

[141] Y. Li, K. Bontcheva, and H. Cunningham. Adapting svm for data sparseness and imbalance : A case study in information extraction. *Natural Language Engineering*, 15(02) :241–271, 2009.

[142] A. Ligeza. A direct approach to planning 2-d collision-free paths for robot. In *Preprints of the IFAC/IFIP/IMACS Sympasium on Theory of Robots*, pages 229–233, 1986.

[143] A. Ligeza. Characteristic functions. an approach to logical reasoning incorporating explicit time representation. Technical Report 39, the Institute of Automatics AGH of Cracow, 1994.

[144] G. Ligozat. Weak representations of interval algebras. In *Proceedings of the 8th National Conference on Artificial Intelligence, AAAI'90*, pages 715–720. AAAI Press, 1990.

[145] G. Ligozat. On generalised interval calculi. In *Proceedings of the 9th National Conference on Artificial Intelligence, AAAI'91*, pages 234–240. AAAI Press, 1991.

[146] M. Lux and S. Chatzichristofis. Lire : lucene image retrieval : an extensible java cbir library. pages 1085–1088, 2008.

[147] A. Mackworth. Constraint satisfaction. In S. Shapiro, editor, *Encyclopedia of Artificial Intelligence*. Wiley Interscience, 1987.

[148] R. Maddux. Relation algebras for reasoning about time and space. In M. Nivat, C. Rattray, T. Rus, G. Scollo, editor, *Algebraic Methodology and Software Technology, Enschede 1993*, Workshops in Computing Series, pages 27–44. Springer-Verlag, 1994.

[149] B. Mandelbrot. Information theory and psycholinguistics : A theory of word frequencies. *Readings in mathematical social sciences*, pages 350–368, 1966.

[150] Z. Manna and A. Pnueli. *The temporal logic of reactive and concurrent systems : Specification*. Springer Verlag, 1992.

[151] Z. Manna and A. Pnuelli. Verification of concurrent programs, part1 : The temporal framework. Technical Report STAN-CS-81-836, Stanford Univresity, 1981.

[152] J. Mari and F. Ber. Temporal and spatial data mining with second-order hidden markov models. *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, 10(5) :406–414, 2006.

[153] P. Marquis. *Contribution à l'étude des méthode de construction d'hypothèses en intelligence artificielle*. PhD thesis, Université de Nancy I, 1991.

[154] J. Martin. The truth, the whole truth, and nothing but the truth. *Artificial Intelligence, Special Issue*, 28 :7–25, 1990.

[155] I. Mau, K. Hornsby, and I. Bishop. Modeling geospatial events and impacts through qualitative change. *LECTURE NOTES IN COMPUTER SCIENCE*, 4387 :156, 2007.

[156] B. McBride. The resource description framework (rdf) and its vocabulary description language rdfs. *Handbook on Ontologies*, pages 51–66, 2004.

[157] A. McCallum, D. Freitag, and F. Pereira. Maximum entropy markov models for information extraction and segmentation. pages 591–598, 2000.

[158] J. McCarthy and P. Hayes. Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, 4 :463–502, 1969.

[159] D. McDermott. A temporal logic for reasoning about actions and plans. *Cognitive Science*, 6 :101–155, 1982.

[160] D. McDermott. A temporal logic for reasoning about processes and plans. *Cognitive science*, 6(2) :101–155, 1982.

[161] D. McGuinness, F. Van Harmelen, et al. Owl web ontology language overview. *W3C recommendation*, 10 :2004–03, 2004.

[162] I. Meiri. Faster constraint satisfaction algorithms for temporal reasoning. Technical Report R-151, UCLA Cognitive System Lab, 1990.

[163] I. Meiri. Combining qualitative and quantitative constraints in temporal reasoning. In *Proceedings of the 9th National Conference on Artificial Intelligence, AAAI'91*, pages 260–267. AAAI Press, 1991.

[164] F. Menczer, G. Pant, P. Srinivasan, and M. Ruiz. Evaluating topic-driven web crawlers. pages 241–249, 2001.

[165] A. Miles and S. Bechhofer. Skos simple knowledge organization system reference. 2008.

[166] H. Miller and Y. Wu. Gis software for measuring space-time accessibility in transportation planning and analysis. *GeoInformatica*, 4(2) :141–159, 2000.

[167] A. Miron, J. Gensel, M. Villanova-Oliver, and H. Martin. Towards the geo-spatial querying of the semantic web with ontoast. *Web and Wireless Geographical Information Systems : 7th International Symposium, W2GIS 2007, Cardiff, UK, November 28-29, 2007, Proceedings*, 2007.

[168] A. Mokkedem and D. Méry. On using temporal logic for refinement and compositional verification of concurrent systems. *Theoretical Computer Science*, 1(140) :95–138, 1995.

[169] R. Morris and L. Khatib. An interval-based temporal relational calculus for events with gaps. *Journal of Experimental and Theoretical Artificial Intelligence*, 3 :87–107, 1991.

[170] R. Morris, W. Shoaff, and L. Khatib. Path consistency in networks of non-convex intervals. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence, IJCAI'93*, pages 655–660. Morgan Kaufmann, 1993.

[171] P. Moulin. The role of information theory in watermarking and its application to image watermarking. *Signal Processing*, 81(6) :1121–1139, 2001.

[172] P. Muller and V. Dugat. *Raisonnements sur l'espace et le temps*, chapter Représentations en logique classique. Hermès, 2007.

[173] M. Najork and J. Wiener. Breadth-first crawling yields high-quality pages. pages 114–118, 2001.

[174] G. Nelson and D. Oppen. Simplifications by cooperating decision procedures. *ACM Transactions on Programming Languages and Systems*, 1(2) :245–257, 1979.

[175] S. Nirenburg, S. Beale, and M. McShane. Evaluating the performance of the ontosem semantic analyzer. 2004.

[176] S. Nunes, C. Ribeiro, and G. David. Use of temporal expressions in web search. *LNCS*, 2008.

[177] S. Overell and S. Rüger. Geographic co-occurrence as a tool for gir. pages 71–76, 2007.

[178] D. Papadias and Y. Theodoridis. Spatial relations, minimum bounding rectangles, and spatial data structures. *IJGIS*, 11(2) :111–138, 1997.

[179] Y. Papakonstantinou, H. Garcia-Molina, and J. Widom. Object exchange across heterogeneous information sources. pages 251–260, 2002.

[180] V. Pavlovic, J. Rehg, T. Cham, and K. Murphy. A dynamic bayesian network approach to figure tracking usinglearned dynamic models. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, 1999.

[181] W. Pedrycz. *Fuzzy Control and Fuzzy Systems*. Research Studies Press Ltd., Taunton, Somerset, England and Tohn Wiley and Sons Inc., 1993.

[182] D. Peuquet. Making space for time : Issues in space-time data representation. *GeoInformatica*, 5(1) :11–32, 2001.

[183] G. M. Provan. Solving diagnostic problems using extended truth maintenance systems. In *Proceedings of 8th European Conference on Artificial Intelligence*, pages 547–553, 1988.

[184] W. Quine. *From a Logical Point of View*. Harper and Row, 1961.

[185] W. Quine. *Ontological Relativity and Other Essays*. Columbia University Press, 1969.

[186] W. Quine. *Philosophy of Logic*. Prentice-Hall, 1970.

[187] S. Rajbhandari, F. Andres, M. Naito, and V. Wuwongse. Semantic-augmented support in spatial-temporal multimedia blog management. *LNCS*, 2007.

[188] A. Reinefeld and P. Ladkin. Fast solution of large interval constraint networks. In J. Glasgow and R. Hedley, editors, *Proceedings of the 9th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI'92*, pages 156–162. Morgan Kaufmann, 1992.

[189] R. Reiter and J. de Kleer. Foundations of assumption based truth maintenance system. In *Proceedings of the 7th National Conference on AI, AAAI'87*, pages 183–188, 1987.

[190] P. Revesz and S. Wu. Spatiotemporal reasoning about epidemiological data. *Artificial Intelligence In Medicine*, 38(2) :157–170, 2006.

[191] R. Robison, L. Henkin, P. Suppes, and A. Tarski. Binary relations as primitive notions in elementary geometry. *The Axiomatic Method, with Special Reference to Geometry and Physics*, pages 30–52, 1959.

[192] M. Runardotter, H. Quisbert, J. Nilsson, A. H
"agerfors, and A. Mirijamdotter. The information life cycle issues in long-term digital preservation. *Arkiv, samh*
"alle och forskning, 2006.

[193] A. Saval. Agate : Advanced geographic alert tool for emergencies, 2007.

[194] A. Saval, M. Bouzid, E. Bondu, and S. Brunessaux. Risk detection and situation awarness : From anyone to everyone. In *S4 Conference Emergence in Geographical Space*, November 2009.

[195] A. Saval, M. Bouzid, and S. Brunessaux. A semantic extension for event modelisation. pages 139–146, 2009.

[196] A. Saval, M. Bouzid, and S. Brunessaux. Vers une modélisation spatiale et sémantique pour l'interprétation des risques. *Rochebrune*, 2009.

[197] A. Saval and Y. Mombrun. Agate : information gathering for risk monitoring. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, page 842. ACM, 2009.

[198] S. Schockaert and M. De Cock. Temporal reasoning about fuzzy intervals, 2008.

[199] R. Schrag, J. Carciofini, and M. Boddy. β-TMM manual. Technical Report CS-R92-012, Honeywell Corp. Systems Research Center, Minneapolis, MN, 1992.

[200] B. Selman and H. Levesque. Abductive and default reasonning : A computational core. In *Proceedings of the 8th National Joint Conference on Artificial Intelligence, AAAI'90*, pages 343–348. AAAI Press, 1990.

[201] C. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 1948.

[202] C. Shannon. Communication theory of secrecy systems. *Journal, vol*, 28(4) :656–715, 1949.

[203] Y. Shoham. *Reasoning About Change : Time and Causation from the Standpoint of Artificial Intelligence*. MIT Press, 1987.

[204] Y. Shoham. Temporal logics in AI : Semantical and ontological considerations. *Artificial Intelligence*, 33(1) :89–104, September 1987.

[205] R. Solomonoff. A preliminary report on a general theory of inductive inference. *Report ZTB-135, Zator Co., Cambridge, MA*, 1960.

[206] J. Stell and M. Worboys. A theory of change for attributed spatial entities. pages 308–319, 2008.

[207] W. Szmielew, L. Henkin, P. Suppes, and A. Tarski. Some mathematical problems concerning elementary hyperbolic geometry. *The Axiomatic Method, with Special Reference to Geometry and Physics*, pages 30–52, 1959.

[208] A. Tarski, L. Henkin, and P. Suppes. What is elementary geometry ? *The Axiomatic Method, with Special Reference to Geometry and Physics*, pages 16–29, 1959.

[209] H. Tolba, F. Charpillet, and J.-P. Haton. Representing and propagating constraints in temporal reasoning. In *Proceedings of the 3th IEEE International Conference on Tools with Artificial Intelligence*, pages 181–184. IEEE Society Press, 1991.

[210] A. Vailaya, A. Jain, and H. Zhang. On image classification : City images vs. landscapes. *Pattern Recognition*, 31(12) :1921–1935, 1998.

[211] E. Valencia. *Outils de topologie algébrique pour la gestion de l'hétérogénéité sémantique entre agents dialogiques*. PhD thesis.

[212] E. Valencia and J. Sansonnet. Model for dialogue between informational agents. *Progress in Artificial Intelligence : 11th Portuguese Conference on Artificial Intelligence, Epia 2003, Beja, Portugal, December 4-7, 2003 : Proceedings*, 2003.

[213] M. Van Assem, V. Malaisé, A. Miles, and G. Schreiber. A method to convert thesauri to skos. *The Semantic Web : Research and Applications*, pages 95–109, 2006.

[214] P. van Beek. Approximation algorithms for temporal reasoning. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence, IJCAI'89*, pages 1291–1296. Morgan Kaufmann, 1989.

[215] J. van Benthem. *The Logic of Time*. D. Reidel, second edition, 1992.

[216] C. Vandeloise. *L'espace en français : sémantique des prépositions spatiales*. Seuil, 1986.

[217] B. Vatant and M. Wick. Geonames ontology, 2007.

[218] Y. Venema. *Many-Dimensional Modal Logic*. PhD thesis, Universiteit van Amsterdam, The Netherlands, 1992.

[219] S. Vere. Planning in time : Windows, durations for activities and goals. In *IEEE actions on Patten Analysis ans Machine Intelligence*, pages 246–267. PAMI, 1983.

[220] S. Vere. Temporal scope of assertions and windows cutoff. In *Proceedings of the 9th International Conference on Artificial Intelligence, IJCAI'85*, pages 1055–1059. Morgan Kaufmann, 1985.

[221] M. Vilian. A system for reasoning about time. In *Proceedings of the 2th National Joint Conference on Artificial Intelligence, AAAI'82*, pages 197–201. AAAI Press, 1982.

[222] M. Vilian and H. Kautz. Constraint propagation algorithms for temporal reasoning. In *Proceedings of the 6th National Conference on Artificial Intelligence, AAAI'86*, pages 377–382. AAAI Press, 1986.

[223] S. Wagon. *The Banach-Tarski Paradox*. Cambridge University Press, 1993.

[224] J. Wallgrun, L. Frommberger, D. Wolter, F. Dylla, and C. Freksa. Qualitative spatial representation and reasoning in the sparq-toolbox. *LECTURE NOTES IN COMPUTER SCIENCE*, 4387 :39, 2007.

[225] B. Williams. Doing time : putting qualitative reasoning on firmer ground. In *Proceedings of the 5th National Conference on AI, AAAI'86*, pages 105–112, 1986.

[226] M. Worboys and K. Hornsby. From objects to events : Gem, the geospatial event model. *LNCS*, 2004.

[227] R. Yangarber, R. Grishman, P. Tapanainen, and S. Huttunen. Automatic acquisition of domain knowledge for information extraction. page 946, 2000.

[228] A. Yao. Theory and application of trapdoor functions. pages 80–91, 1982.

[229] M. Zweben, E. Davis, E. Drascher, and M. D. M. Eskey. Learning to improve constraint-based scheduling. *Artificial Intelligence*, 58 :271–296, December 1992.

[230] Wolfram|Alpha *http ://www.wolframalpha.com/*

[231] OW2 *http ://www.ow2.org/*

[232] WebLab *http ://www.weblab-project.org/*

[233] EADS *http ://www.eads.com/*

[234] SAIMSI *http ://www.systematic-paris-region.org/en/projets/saimsi*

Résumé

La popularité des réseaux sociaux et les nouvelles formes de communication a entraîné l'apparition de nouvelles sources d'informations qu'il convient d'étudier. N'importe qui est en mesure de publier et de mettre en avant les informations qui l'intéresse. Aujourd'hui, ces comportements apparaissent comme autant de moyens de suivre l'évolution d'un sujet d'intérêt : par exemple la grippe H1N1. Cependant, le traitement automatique de ces informations reste encore à améliorer pour pouvoir définir sémantiquement un sujet d'intérêt (les implications du Tsunami de Myanmar). Cette thèse propose une extension sémantique de la modélisation d'événements dans le temps et dans l'espace pour représenter l'évolution de ces sujets d'intérêt. Ce rapport explique comment utiliser le formalisme introduit pour définir des méthodes de raisonnement sur une base de connaissance structurée afin d'améliorer la représentation de situation par la découverte de relations dans ces informations.

Abstract

The popularity of social networks and new forms of communication has led to the emergence of new sources of information that should be studied. Anyone is able to publish and highlight information of interest. Today, these behaviors appear as ways to track a topic of interest : eg H1N1. However, the automatic processing of such information needs to be improved in order to define semantically a topic of interest (the implications of the Tsunami in Myanmar). This thesis propose a semantic extension of the modeling of events in time and space to represent the evolution of these topics of interest. This report explains how to use the introduced formalism to define the methods of reasoning on a knowledge base structured to improve the representation of the situation by discovering relationships in this information.