Chapitre 6

Recherche d'informations géographiques sur Internet

Recherche d'informations géographiques sur Internet

- 6.1 – Principes
- 6.2 – SpatialML
- 6.3 – Système Dési
- 6.4 – Conclusions

6.1 – Principes

- Je veux tout savoir sur un lieu :
- Epoque 1 : recherche dans la BD géo
- Epoque 2 : recherche dans tout Internet

Types de recherche géographique sur Internet

- Tout connaître sur un endroit :
  - Maintenant
  - Dans le passé
- Meilleur façon d'aller de A à B (tout système de transport confondu)
- Transformer un texte en carte
  - Météorologie
  - Récits d'explorateurs
- Etc.

Difficultés

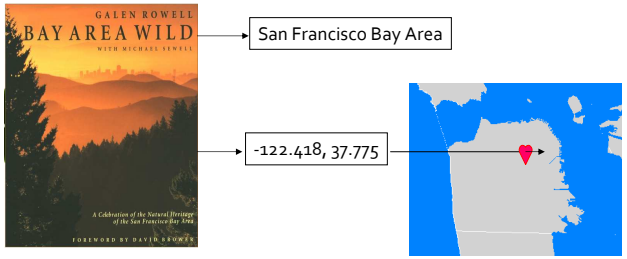
- Tâche gigantesque (indexation spatiale des documents)
- Contrôle de qualité
- Algorithmes pour répondre aux questions
- Algorithmes se donnant une limite de temps
- Confidentialité de localisation

Geographic Information Retrieval (GIR)

- GIR est concerné par la recherche des ressources informationnelles géographiques qui peuvent être pertinentes pour la zone géographique de requête

Geocoding

- Exemple



Préalables

- Analyser les textes existants
  - Extraction des noms de lieux (dates)
    - Geoparsing
  - Codification
    - Geocoding
- Approche spatiale des documents
- Approche sémantique des documents

Approche spatiale

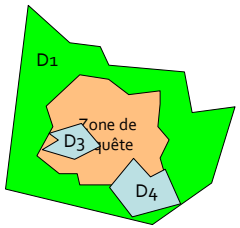
- Usage des représentations spatiales et des relations spatiales
- Approche spatiale
  - Quantitative, basée sur les propriétés géométriques des objets
  - Qualitative, basée sur les propriétés non-géométriques des objets

Similarité spatiale

- Indicateur de pertinence
- Adéquation entre
  - Lieu (décrit dans un document)
  - Lieu (décrit dans une requête)
- Méthode
  - Degré de recouvrement

Recouvrement

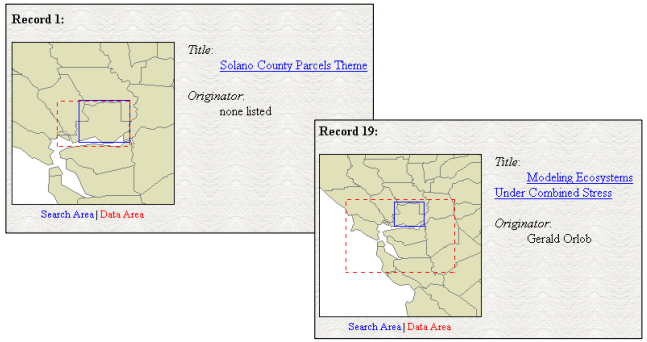
- Retrouver les documents tels que leurs objets géographiques soient en recouvrement avec la zone de requête
- Y inclure les objets totalement inclus, partiellement inclus, totalement recouverts dans la zone de requête
- Relations topologiques, pas de raffinement métriques



Degré de recouvrement

- Calcul de la surface en intersection
- % de recouvrement

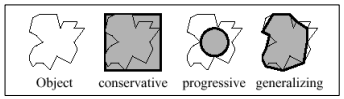
Exemple



http://calsip.regis.berkeley.edu/pattyf/mapserver/cheshire2/cheshire_init.html

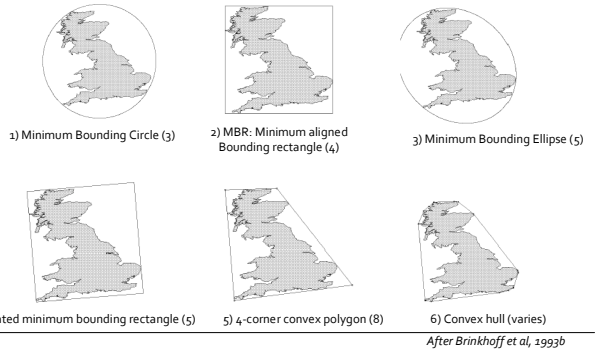
Calcul d'intersection

- Calcul exact
- Approximations
- Système d'accélération (généralisation)



- Opérations plus facile sur des polygones connexes et convexes

Autres approximations possibles



After Brinkhoff et al, 1993b

Méthodes de calcul de la similarité

Reference	Formula
Hill, 1990[10]	$Range = 2\frac{Q}{Q+C}$
Walker et al, 1992[19]	$Range = MIN\left(\frac{Q}{Q}, \frac{Q}{C}\right)$
Beard and Sharma, 1997[3]	Case 1: Q contains C $Range = \frac{C}{Q}$ Case 2: Q and C overlap $Range = \frac{O\%}{(1-O\%)+100}$ Case 3: Q contained in C $Range = \frac{Q}{C}$
Where: Q = area of query region C = area of candidate GIO O = area of overlap for G, C	Range (for all): 0 = no similarity 1 = identical

Méthode de Larson et Frontiera

- Probabilité de pertinence est basée sur une régression logistique afin de déterminer les coefficients (c_k)

$$P(R \mid Q,D) = c_0 + \sum_{i=1}^{m} c_i X_i$$

- Les X_i étant définis ainsi :

Définition des X_i

- X_1 = aire de superposition (zone de requête, zone du document) / aire de la zone de requête
- X_2 = aire de superposition (zone de requête, zone du document) / aire de la zone du document
- Où X_i sont comprises entre 0 et 1

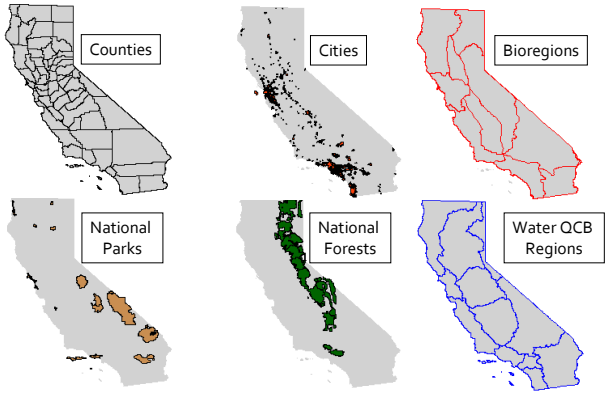
Collection de test

- California Environmental Information Catalog (CEIC)
- http://ceres.ca.gov/catalog.
- Environ 2500 documents sélectionnés dans une collection (août 2003) d'environ 4000.

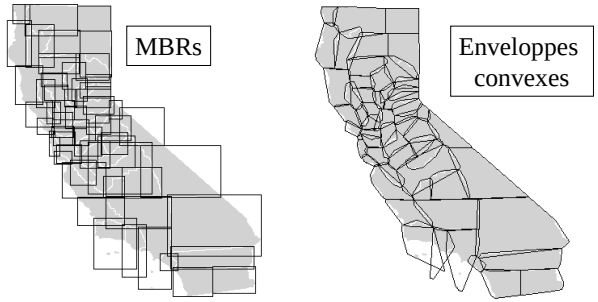
Vue d'ensemble de ces documents

- 2554 documents avec métadonnées indexés par 322 régions géographiques uniques (représentées comme MBRs) and associés à des noms de lieux
  - 2072 documents (81%) indexés par 141 toponymes uniques en CA
    - 881 documents avec 42 noms de comtés (sur 46)
    - 427 documents avec 76 noms de villes (sur 120)
    - 179 documents avec 8 noms de bioregions (sur 9)
    - 3 documents avec 2 noms de parcs nationaux (sur 5)
    - 309 documents avec 11 noms de forêts domaniales (sur 11)
    - 3 documents de 1 agence de l'eau (sur 1)
    - 270 documents sur 1 état (CA)
  - 482 documents (19%) indexés par 179 autres zones
    - 12% représentant des régions à l'intérieur de la Californie
    - 88% (158 sur 179) régions côtières et îles

Toponymes en Californie



Approximations géométriques



Résultats du calcul

Approximation	Logistic Regression Model Fitted on the Training Data
MBR	$LogO(R\|Q,C) = -5.0402 + (6.5154 * X_1) + (5.7729 * X_2)$
Convex Hull	$LogO(R\|Q,C) = -3.4767 + (7.4536 * X_1) + (5.7569 * X_2)$

Comparaison des méthodes

Seulement avec toponymes indexés

Ranking Method	MBRs	Convex Hulls
Hill, 1990	0.7193	0.8097
Walker et al., 1992	0.7025	0.8006
Beard & Sharma, 1997	0.7094	0.8116
Logistic Regression	0.9389	0.9973

Sur l'ensemble des données

Ranking Method	MBRs	Convex Hulls
Hill, 1990	0.6722	0.7936
Walker et al., 1992	0.6509	0.7810
Beard & Sharma, 1997	0.6523	0.7778
Logistic Regression	0.8141	0.9099

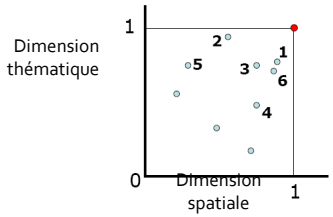
La méthode de Larson et Frontiera donne les meilleurs résultats

Expansion de requêtes et indexation spatiale

- Expansion de requêtes signifie que si un utilisateur recherche des documents sur Lyon, il faut ajouter Villeurbanne, Caluire, etc. à la question.
- Ainsi besoin de connaître la topologie, les lieux voisins et leurs relations spatiales.
- On doit typiquement utiliser des dictionnaires géographiques. Si un utilisateur demande des documents à propos des "châteaux à Zurich", un index spatial réduit le numéro de documents seulement à cette ville et à son voisinage.

Classification

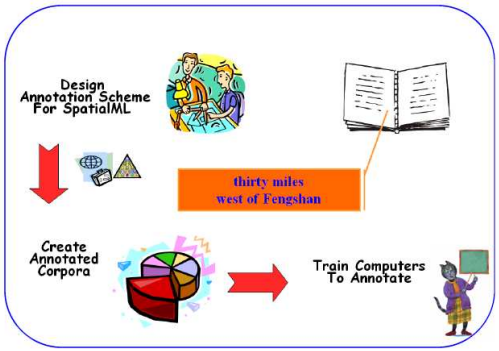
- Un système de recherche des informations ramène des documents avec un score
- En GIR les scores ont typiquement, deux dimensions (thématique et spatiale)



6.2 – SpatialML

- La recherche d'information est rendue difficile à cause de l'absence de formalisation
- Création de SpatialML comme langage à balises (XML) pour représenter les lieux et leurs relations dans les textes en langues naturelles
- Système d'annotation
- MITRE corporation

Exemple



Codage

a <PLACE id="1" type="FAC" form="NOM">building</PLACE>

<SIGNAL id="2">5 miles</SIGNAL>

<SIGNAL id="3">east</SIGNAL>

of <PLACE id="4" type="PPL" country="TW" form="NAM"
latLong="22°37'N 120° 21'E">Fengshan</PLACE>

<PATH id="5" source="4" destination="1" distance="5:mi"
direction="E" signals="2 3" frame="EXTRINSIC"/>

Exemple multilingue

I live in a [town] some [50 miles] [south] of [Salzburg] in the central [Austrian] [Alps].

جبال الالب النمسا و سالزبرج فى وسط جنوب خمسين ميل مدينة تبعد حوالى أنا أسكن فى

<PLACE type="PPL" id=1 form="NOM">مدينة</PLACE>
<SIGNAL id=2>خمسين ميل </SIGNAL>
<SIGNAL id=3>جنوب /SIGNAL>
<PLACE id=4 type="PPLA" country="AT" form="NAM">سالزبرج </PLACE>
<PLACE id=5 type="COUNTRY" country="AT" mod=C>النمسا <PLACE>
<PLACE id =6 type="MTS">جبال الالب </PLACE>
<PATH id=7 distance="50:mi" direction=S source= 4 destination=1 signals="2 3"/>
<LINK id=8 source=1 target=6 linkType="IN"/>

나는 [오스트리아] [알프스] 중심의 [잘츠부르크] [남쪽]에서 [50마일] 거리의 마을에 산다

<PLACE type="PPL" id=1 form="NOM" ctv="TOWN">마을</PLACE>
<SIGNAL id=2>50 마일</SIGNAL>
<SIGNAL id=3>남쪽</SIGNAL>
<PLACE id=4 type="PPLA" country="AT" form="NAM">잘츠부르크</PLACE>
<PLACE id=5 type="COUNTRY" country="AT" mod="C">오스트리아</PLACE>
<PLACE id=6 type="MTS" >알프스</PLACE>
<PATH id=7 distance="50:mi" direction=S source= 4 destination=1 signals="2 3"/>
<LINK id=8 source=1 target=6 linkType="IN"/>

Types de lieux

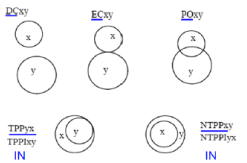
BODYOFWATER	River, stream, ocean, sea, lake, canal, aqueduct, geyser, etc.
CELESTIAL	sun, moon, Jupiter, Gemini, etc.
CIVIL	Political Region or Administrative Area, usually sub-national, e.g. State, Province, certain instances of towns and cities.
CONTINENT	Denotes a continent, including ancient ones. See Table 1.
COUNTRY	Denotes a country, including ancient ones. See Table 1.
FAC	Facility, usually a catchall category for restaurants, churches, schools, ice-cream parlors, bowling alleys, you name it!
GRID	A grid reference indication of the location, e.g., MGRS (Military Grid Reference System)
LATLONG	A latitude/longitude indication of the location
MTN	Mountain
MTS	Range of mountains
POSTALCODE	Zipcodes, postcodes, pincodes etc.
POSTBOX	P. O. Box segments of addresses
PPL	Populated Place (usually conceived of as a point), other than PPLA or PPLC
PPLA	Capital of a first-order administrative division, e.g., a state capital.
PPLC	Capital of a country
RGN	Region other than Political/Administrative Region
ROAD	Street, road, highway, etc.
STATE	A first-order administrative division within a country, e.g., state, province, gubernia, territory, etc.
UTM	A Universal Transverse Mercator (UTM) format indication of the location
VEHICLE	Car, truck, train, etc.

Relations entre les lieux

a <PLACE id="1" form="NOM" type="FAC">school</PLACE> in
<PLACE id="2" form="NAM" type="PPL" latLong="39.952°N
75.164°W">Philadelphia</PLACE>

<LINK source=1 target=2 linkType="IN"/>

RCC8 Spatial Calculus*



*http://www.irit.fr/~Philippe.Muller/Publis/ci02.pdf

LinkType	Example
IN (tangential and non-tangential proper parts)	[Paris], [Texas]
EC (extended connection)	the border between [Lebanon] and [Israel]
NR (near)	visited [Belmont], near [San Mateo]
DC (discrete connection)	the [well] outside the [house]
PO (partial overlap)	[Russia] and [Asia]
EQ (equality)	[Rochester] and [382044N 0874941W]

Orientations

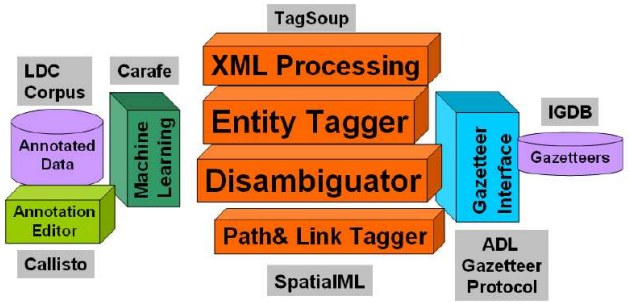
MOD Code	Example
B	the bottom of the [well]
BR	[Burmese] border
C	central [district]
E	eastern [province]
N	[North India]
NEAR	near [Harvard]
S	southern [India]
T	the top of the [mountain]
W	west [Tikrit]

Direction Code	Example
B	[behind] the house
A	[above] the roof
BL	[below] the tree-line
E	[E] of
ESE, WSW, etc.	
F	[in front of] the theater
N	[north] of
S	[south] of
W	[W] of

Tags SpatialML

- **PLACE** : tags qui indiquent le type d'endroit, le répertoire de toponymes utilisé et les coordonnées
- **LINK** tags qui expriment la superposition, la connexion ou autres relations topologiques entre un couple d'endroit
- **RLINK** (RELATIVE-LOCATION-LINK) tags qui expriment l'aspect relatif d'un endroit par rapport à un autre

Architecture

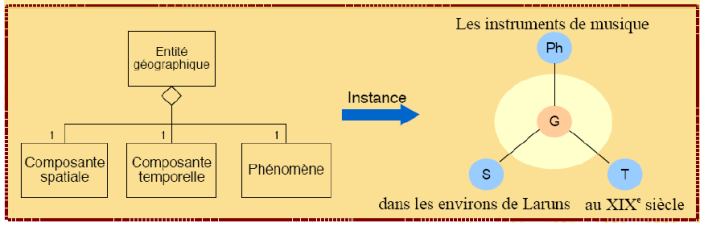


6.3 – Système Dési

- Système d'analyse des textes pour y retrouver les
  - composantes spatiales
  - composantes temporelles
- Développé par l'Université de Pau

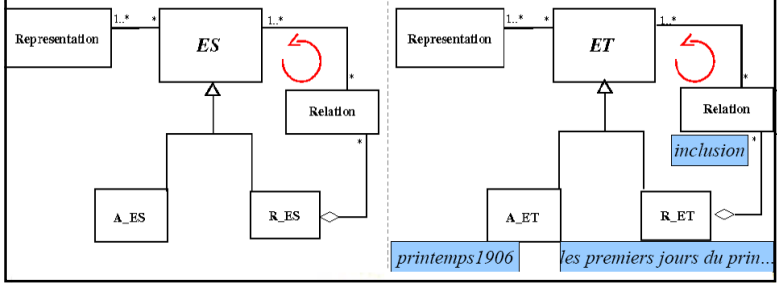
Exemple initial

- Information Géographique
  - Espace, Temps, Phénomène (Thème)
  - Exemple : « Les instruments de musique dans les environs de Laruns [...] aux XIXème siècle »



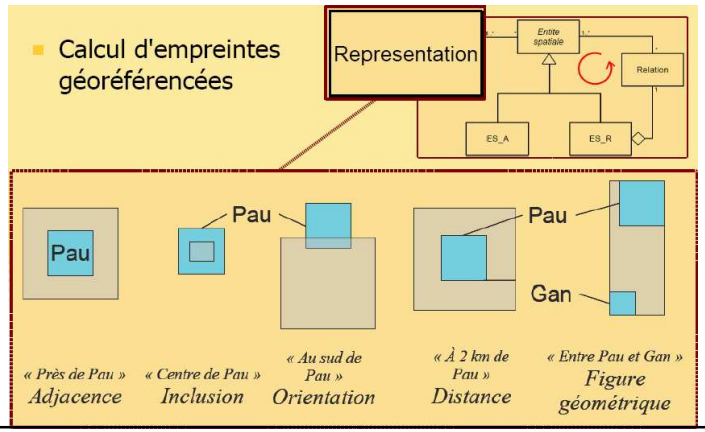
Exemple de texte à analyser

(1) [Vers les premiers jours du printemps 1906] je décidai d'aller à la cabane [à deux heures de Gavarnie]

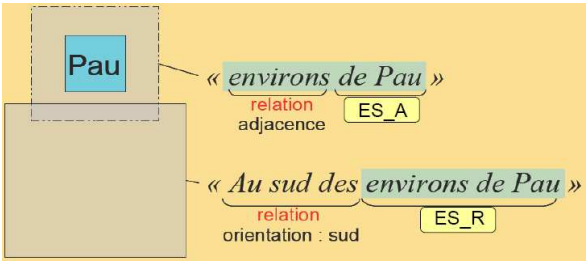


Index spatial

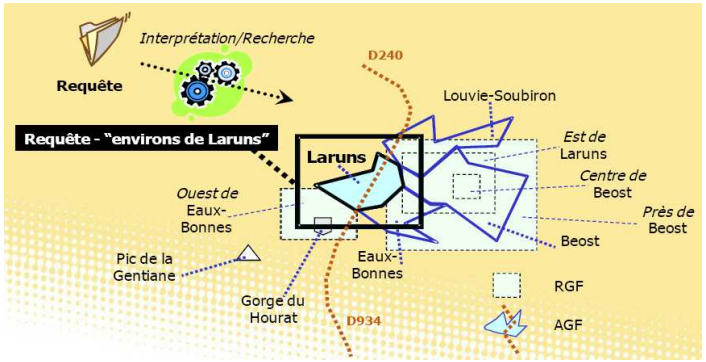
- Calcul d'empreintes géoréférencées



Aspects spatiaux des requêtes



Exemple de requête



6.4 – Conclusions

- Importance de la recherche d'informations géographiques sur Internet
- Nécessité d'annoter les pages-web
- Nécessité de les indexer