Trouver de quoi parle un article sans le comprendre

Click to edit Master subtitle style
Ihab Mallak, Pierre Prade, Mohand Boughanem
IRIT – Université de Toulouse
Equipes SIG - RPDMP

Objectif (1)

- Parcourir un texte sans l'analyser en repérant les mots significatifs
- On ne cherche pas à résumer le texte
- mais plutôt à avoir une idée de son contenu sémantique

Objectif (2)

- Identifier les thèmes d'un texte:
  - Identification des mots « significatifs »
  - À partir ces mots, identifier les phrases « informatives »

- Significatifs, informatives sont des notions graduelles, floues

- D'où l'utilisation des ensembles flous

Total on trial over 1999 French oil disaster
11 Feb [illegible] GMT
Source: Reuters
Alert Me | Printable view | Email this article | RSS [-] Text [+]
By James Mackenzie
PARIS, Feb 12 (Reuters) - A trial into one of France's worst environmental disasters opens on Monday with oil giant Total <TOTF.PA> facing charges over toxic fuel spills that washed ashore following the sinking of the tanker Erika in 1999.
Total is among 15 organisations and individuals charged over the spill that poured 20,000 tonnes of oil into the sea, polluted 400 km (250 miles) of coastline and caused damage valued at up to 1 billion euros ($1.30 billion).
... ...
The French government alone is seeking 153 million euros.
The trial [illegible] in which a multinational will face charges on maritime pollution in France, is expected to last until June at least.
Besides Total and two of its subsidiaries, the ship's Indian captain, its management company, four French maritime officials and the Italian maritime certification company RINA, which classified the ship as safe, are also on trial.
An army of lawyers, some 69 witnesses and interpreters in Italian, English and Hindi will take part in the proceedings in the Tribunal de Grande Instance in Paris.
Critics, including the environmental group Friends of the Earth, which is one of the plaintiffs in the trial, say Total took cynical risks with the ship to meet a tight commercial deadline.
They say international maritime law still needs to be tightened to minimise risks to the environment.

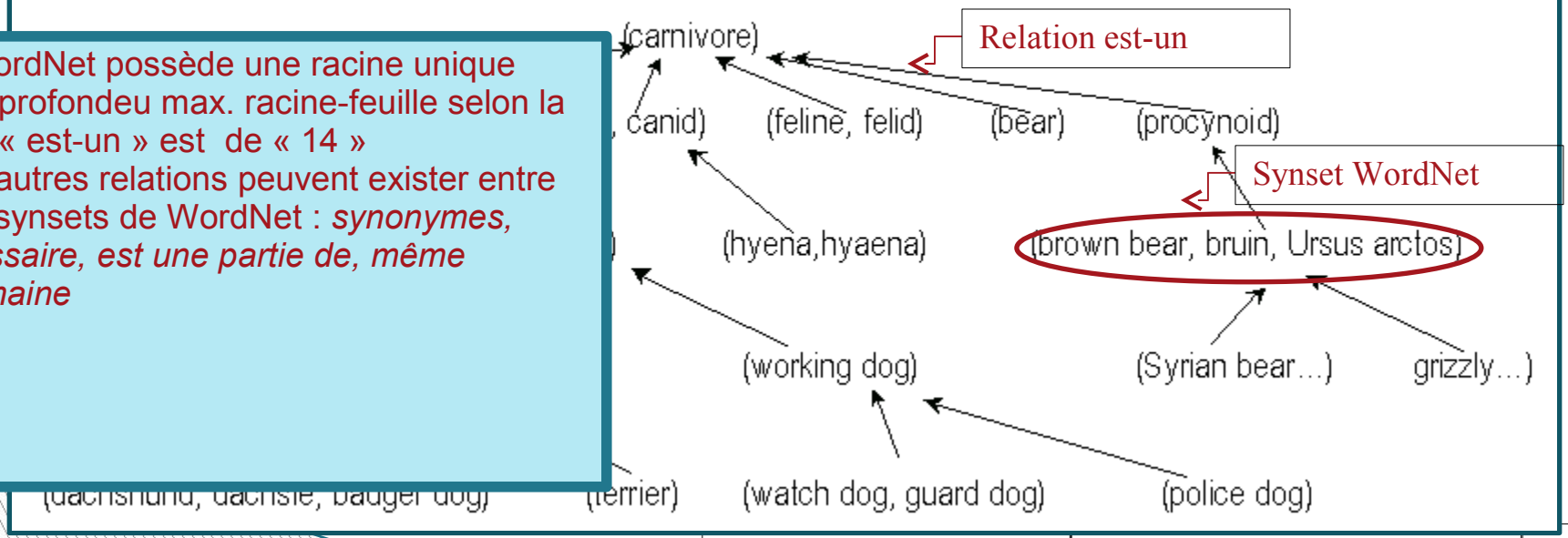
Principes de base en RI

- En RI l'extraction des mots-clés d'un document est basée sur:
  - Des mots simples extraits du document à l'aide d'indices fréquentiels: *tf*idf* …
  - Des concepts extraits à partir des ressources externes: *Dictionnaire, thesaurus, ontologie*

WordNet

- Une base terminologique
- Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise
- Exemple:

Dans l'exemple ci-dessous, est présenté une sous hiérarchie de WordNet correspondant au concept " chien "



- WordNet possède une racine unique
- la profondeu max. racine-feuille selon la rel. « est-un » est de « 14 »
- D'autres relations peuvent exister entre les synsets de WordNet : *synonymes, glossaire, est une partie de, même domaine*

Plan

Approche

1. Représentation du contenu d'un document
1. Extraction des termes représentatifs basée sur des fonctions d'évaluation floues
1. Identification des phrases significatives d'un document

Expérimentation

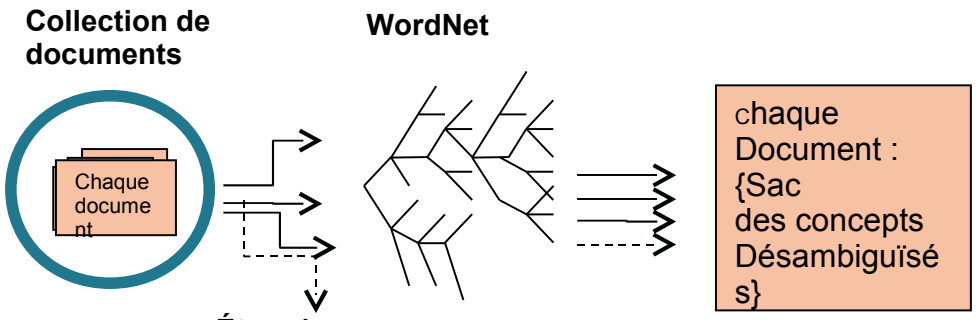
Approche

1- Représentation du contenu du document

La méthode d'extraction est basée sur les synsets de WordNet

Pour chaque document:

- Sélection des termes (substantifs)
- Identification pour chaque terme son synset dans WordNet:
  - En cas de mots ou d'expression polysémiques, nous utilisons une méthode de désambigüisation pour sélectionner le sens le plus plausible
- Clustérisation des termes



Approche

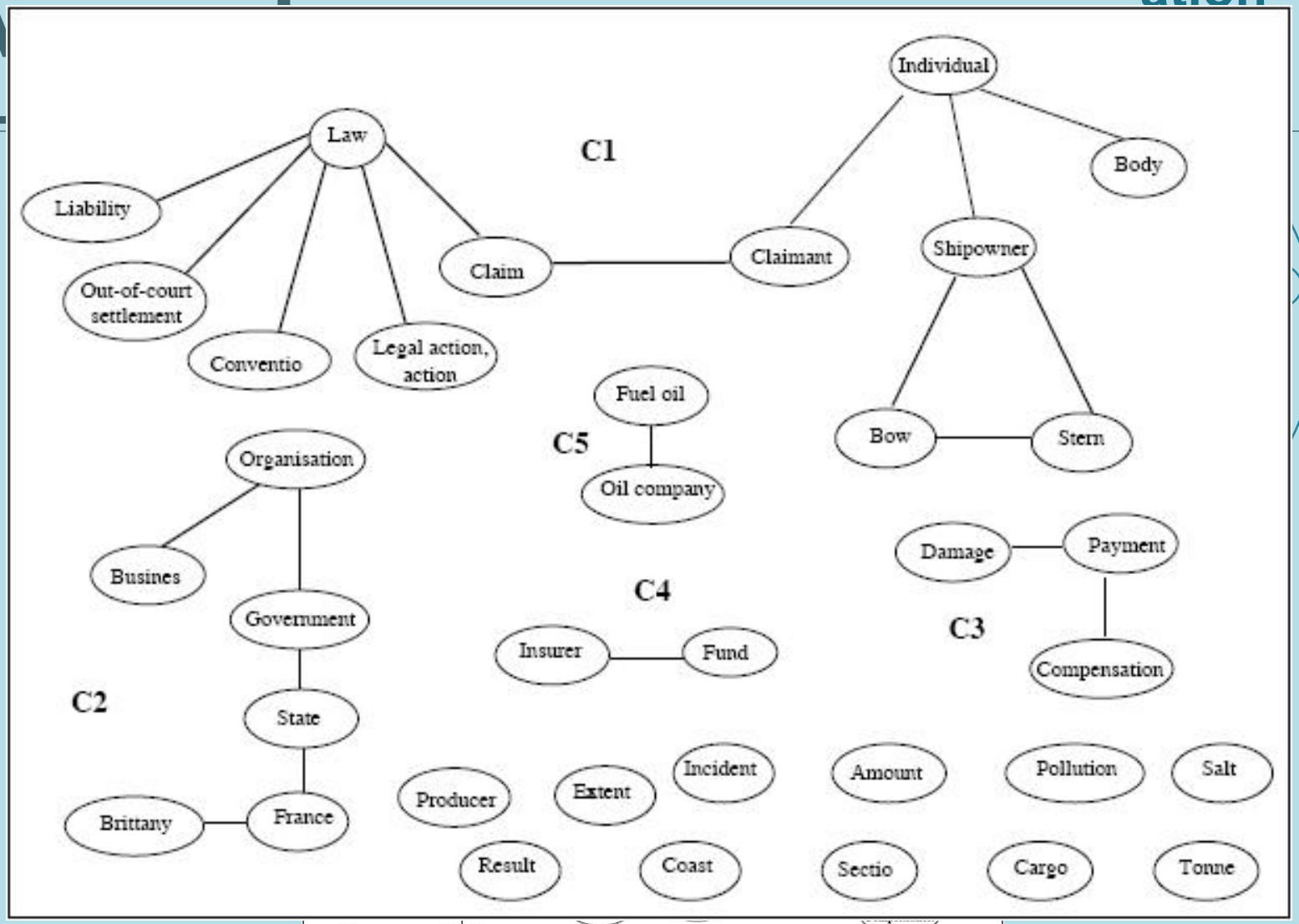
1-Représentation du contenu du document

Clustérisation:

À partir du sac de concepts

- Identifier les relations sémantiques entre ces différents concepts
  (Utilisant les relations WordNet :*synonymes, glossaire, spécialisation, généralisation, est une partie de, même domaine*)
- Grouper les concepts reliés entre eux dans des clusters

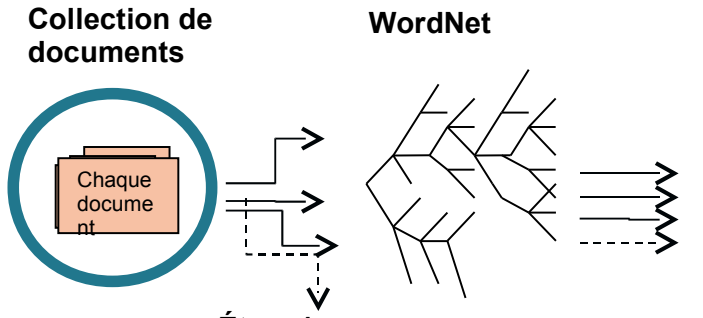
A
1.

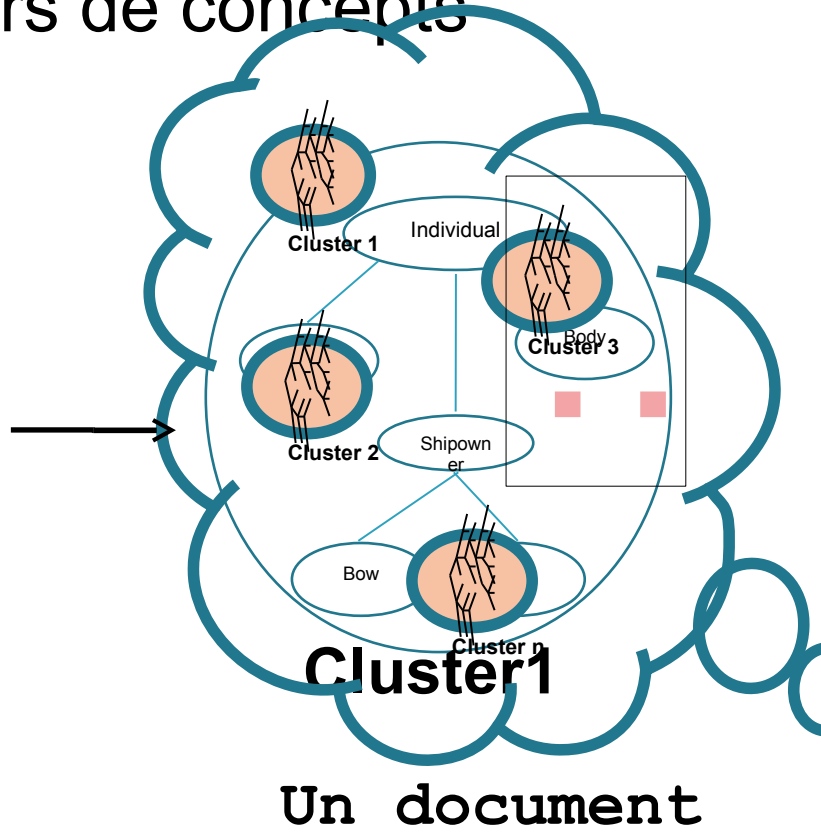

Individual
Law
C1
Body
Liability
Claim
Claimant
Shipowner
Out-of-court settlement
Conventio
Legal action, action
Fuel oil
C5
Bow
Stern
Organisation
Oil company
Damage
Payment
Busines
C4
C3
Government
Insurer
Fund
Compensation
C2
State
Incident
Amount
Pollution
Salt
Producer
Extent
Brittany
France
Result
Coast
Sectio
Cargo
Tonne

Approche

1- Représentation du contenu du document

Après le groupement des concepts le document sera représenté par plusieurs clusters de concepts





Approche

1- Représentation du contenu du document

- Un cluster regroupe des termes qui font référence à une même idée conceptuelle

- Cette idée peut être très présente dans un document même si chacun des termes qui y font référence sont peu fréquents dans un document

- On améliore ce que capture le « tf» (fréquence d'un terme dans un document) calculé sur des mots isolés

Approche

2-Extraction des termes significatifs

À chaque itération de la procédure:

- Identifier les clusters dans lesquels on extrait les termes significatifs
  - Calculer le poids d'un cluster (*somme des fréquences des termes qui le composent*)

  - sélectionner les clusters ayant le poids le plus important *(avec un seuil de tolérance)*

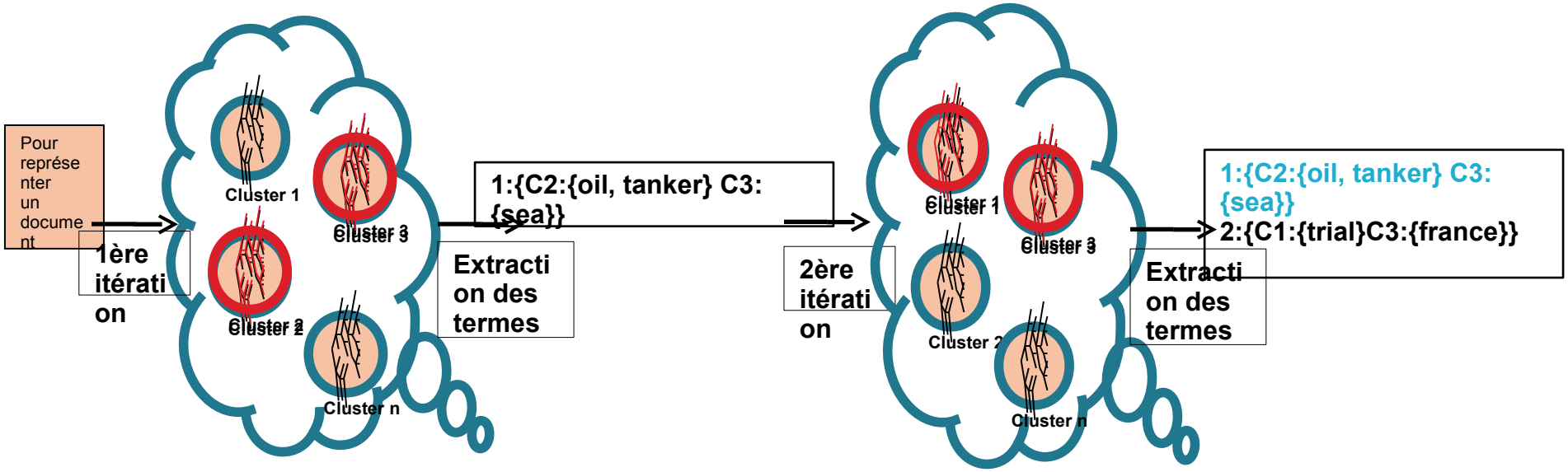
  - Extraire les termes les plus significatifs *(les termes sont ensuite supprimés du cluster)*

- à chaque nouvelle itération on trouve des mots de moins en moins significatifs
- Le test d'arrêt est basé sur le nombre de mots extraits par rapport au nombre total de mots dans le document

Approche

2-Extraction des termes significatifs

Exemple



Approche

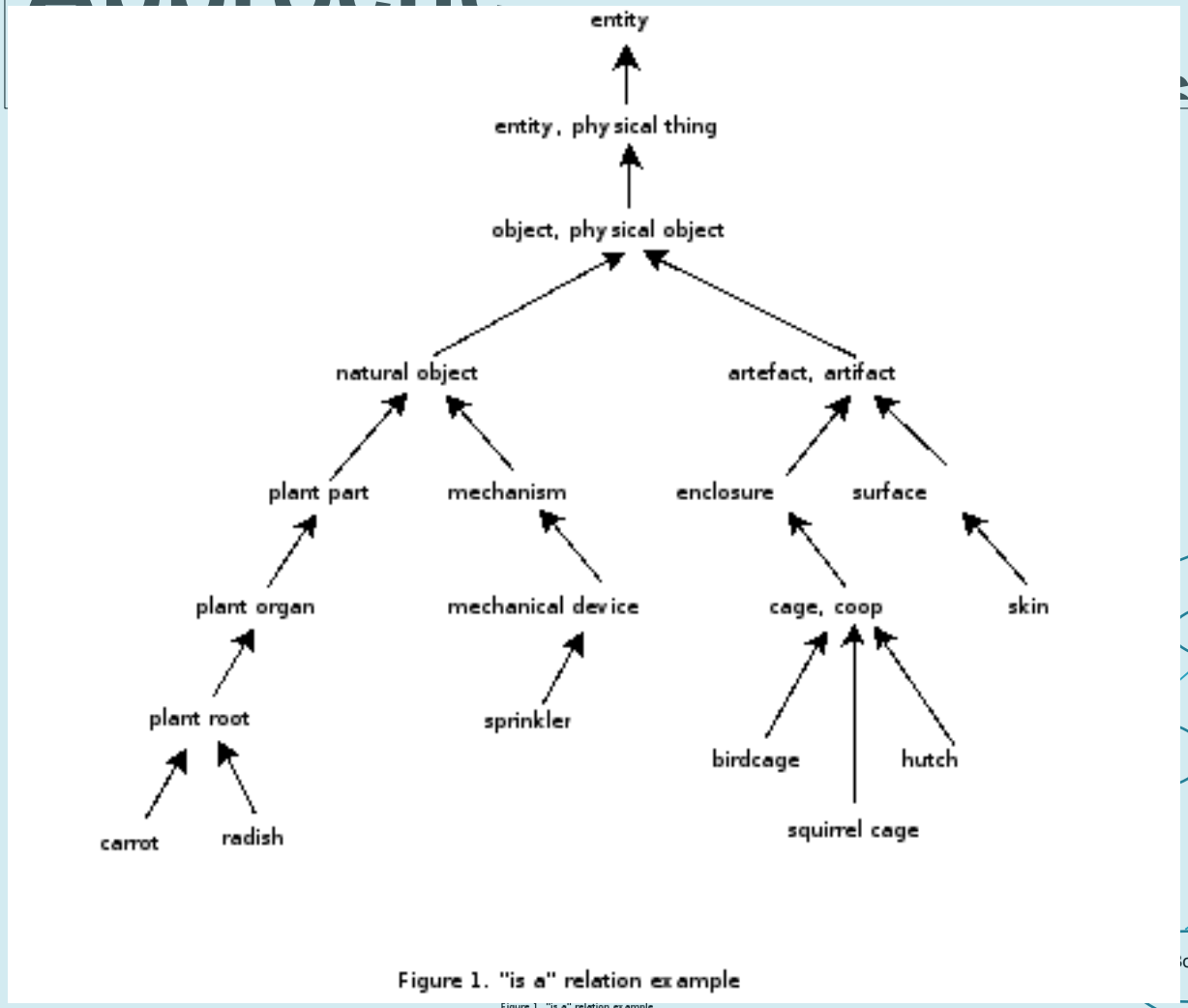
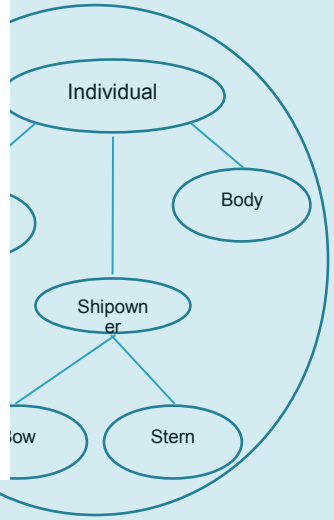


Figure 1. "is a" relation example

Figure 1. "is a" relation example



calcule de spécificité de « plant part » = 4

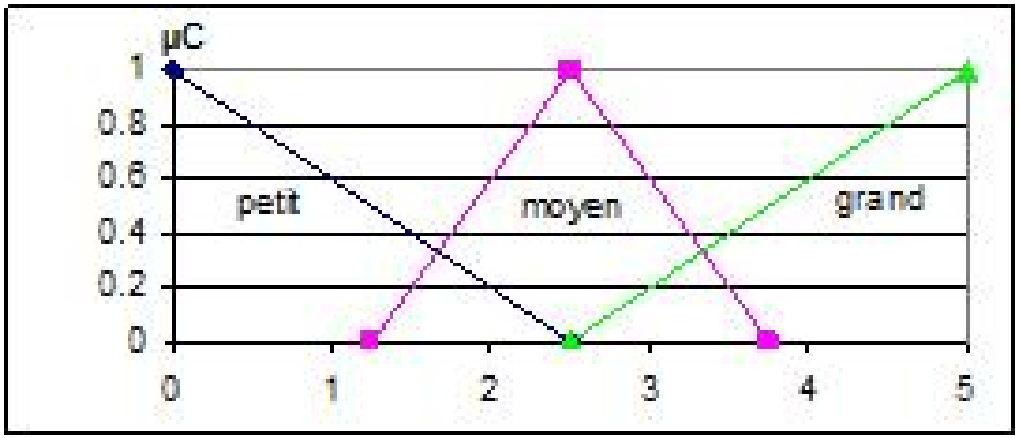
Centralité (Shipowner) =3

Approche

2-Extraction des termes significatifs

- Fct= min(μTF-assez-grand, max(μs-grand,μc-grand))

Ces paramètres sont représentés par des ensembles flous sur des partions floues *'grand','moyen' et 'petit'*



Approche

2-Extraction des termes significatifs

- Comparaison entre la sortie de notre méthode et la fonction standard tf*idf

Fct= min(µTF-assez-grand, max(µs-grand,µc-grand))	tf * idf
1:{C3:{ship} } 2:{C3:{trial, case} } 3:{C3:{charge, damage, tanker} } 4:{C3:{government, company, oil} } 5:{C3:{proceeding, tribunal, faith, individual, contract, article, witness, official, council, friend, certification, crew_member, lawyer, critic, law} }	Focus, paris, faith, election, trial, ten_thousand, world, critic, france, june, euro, fine, proceeding Cargo, week, issue, instance, law, may, alert, km, contract, helicopter, company

- Utilisation de termes significatifs:
  - représentation un document par un groupe des mots représentatifs sur différents niveaux
  - Indexation
  - Représentation des thèmes importants dans un document

Approche

3- Identification des phrases significatives

- *snippets google*

snippets google

Approche

3- Identification des phrases significatives

Idéalement les phrases à sélectionner doivent donner une bonne idée du contenu du document

Ces phrases sont celles:

i. Qui contient le nombre maximum de mots parmi les plus représentatifs
ii. Où les mots provenant d'un grand nombre de clusters
iii. Où les phrases sélectionnées sont de taille courtes

Dans l'expérimentation nous avons une procédure plus simple:

Sélectionner d'abord les phrases contenant une proportion suffisamment élevée de termes significatifs

Approche

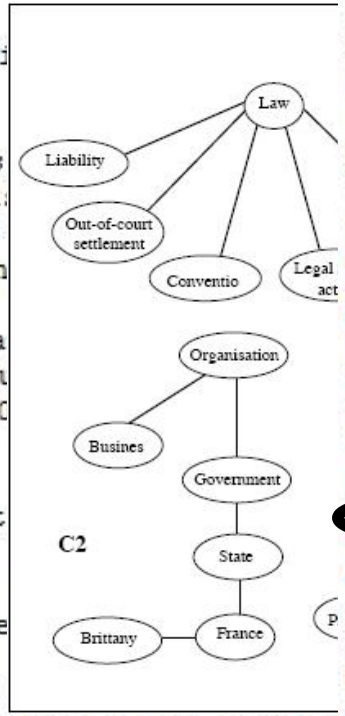
4- Expérimentation

- Collections de 20 documents issus de web portant sur le naufrage de Erika.
- Chaque article traite l'événement d'un point de vue différent

But de cette expérimentation:

- Extraire à partir de ces documents des phrases informatives qui les représentent au mieux
- Comparer ces phrases à des thèmes extraits manuellement par des utilisateurs

Total on trial over 1999 French oil disas
11 Feb 2007 23:19:06 GMT
Source: Reuters
Alert Me | Printable vi
] Text [+]
By James Mackenzie
PARIS, Feb 12 (Reuters
worst environmental di
giant Total <TOTF.PA>
spills that washed ash
tanker Erika in 1999.
Total is among 15 orga
over the spill that pou
sea, damage polluted 4(
caused valued at up to
... ...
The French government
euros.
The trial itself, the
multinational will face
France, is expected to
Besides Total and two
Indian captain, its management compar
maritime officials and the Ita
certification company RINA, which classi
safe, are also on trial.
An army of lawyers, some 69 witnesses
in Italian, English and Hindi will ta
proceedings in the Tribunal de Grande Ins
Critics, including the environmental g
the Earth, which is one of the plaintif
say Total took cynical risks with the
tight contract deadline.
They say international maritime law st
tightened to minimise risks to the enviro



Total on trial over 1999 French oil disaster
11 Feb 2007 23:19:06 GMT
Source: Reuters
Alert Me | Printable view | Email this article | RSS [-
] Text [+]
By James Mackenzie
PARIS, Feb 12 (Reuters) - A trial into one of France's worst environmental disasters opens on Monday with oil giant Total <TOTF.PA> facing charges over toxic fuel spills that washed ashore following the sinking of the tanker Erika in 1999.
Total is among 15 organisations and individuals charged over the spill that poured 20,000 tonnes of oil into the sea, polluted 400 km (250 miles) of coastline and caused damage valued at up to 1 billion euros ($1.30 billion).
... ...
The French government alone is seeking 153 million
The trial itself, the first on such a scale in which a multinational will face charges on maritime pollution in France, is expected to last until June at least.
Besides Total and two of its subsidiaries, the ship's Indian captain, its management company, four French maritime officials and the Italian maritime certification company RINA, which classified the ship as safe, are also on trial.
An army of lawyers, some 69 witnesses and interpreters in Italian, English and Hindi will take part in the proceedings in the Tribunal de Grande Instance in Paris.
Critics, including the environmental group Friends of the Earth, which is one of the plaintiffs in the trial, say Total took cynical risks with the ship to meet a tight contract deadline.
They say international maritime law still needs to be tightened to minimise risks to the environment.

Doc n°	Lignes extraits du texte	Nb de termes dans les lignes extraits	Nb de termes dans le document	Thèmes générale du document extrait manuellement
	* Total guilty over Erika oil spill.			* French oil giant Total responsible for the 1999 sinking of the tanker Erika,

Doc n°	Phrases extraites du texte	Nb de termes dans les lignes extraits	Nb de termes dans le document	Thèmes généraux du document extraits manuellement
20	* Prosecutor wants Total convicted for Erika disaster ** PARIS (Reuters) - French oil giant Total should be convicted of maritime pollution for its role in the sinking of the oil tanker Erika, which provoked one of France's worst environmental disasters, prosecutors said on Monday. - The company denies the charges	48	295	* Prosecutor wants Total convicted for Erika disaster - Erica history and effects seabirds ** Total failed to conduct proper checks before chartering the ageing ship. ** Total had faced pollution and negligence charges as well as complicity in endangering human lives over the incident. * Prosecution convict six other individuals and organizations

Discussion des résultats

- Les résultats obtenus sont de bonne qualité (83.3% des thèmes manuellement extraits ont été identifiés):
  - Cependant quelques termes extraits par la méthode de base ne sont pas présents dans les phrases extraites automatiquement

Conclusion et perspective

Conclusion:

- La méthode proposée permet de présenter une vue sémantique du contenue d'un texte

- elle est plus riche sémantiquement que celle que fournirait une approche purement statique du type *tf*idf*

Perspective:

- Evaluer l'impact des différents paramètres *(fréquence, centralité, spécificité)* sur l'extraction des documents pour répondre à des requête


Questions