UQÀM
Université du Québec à Montréal

DIC9410 Présentation du projet de recherche

Doctorat en Informatique Cognitive

Maintenance d'ontologies de domaine à partir d'analyses textuelles

Yassine Gargouri

Codirecteurs de Recherche :

Bernard Lefebvre & *Jean-Guy Meunier*

11 février 2004

Plan de présentation

- **Introduction**
- **Problématique**
  - État de l'art
  - Composante cognitive
  - Composante informatique
  - Choix théoriques
- **Proposition de solution et méthodologie**
  - Modèle proposé
  - Méthodologie de recherche
  - Méthode de validation des résultats
  - Plan sommaire de la thèse
  - État d'avancement des travaux
- **Conclusions**
  - Contributions originales
  - Obstacles à franchir

Introduction

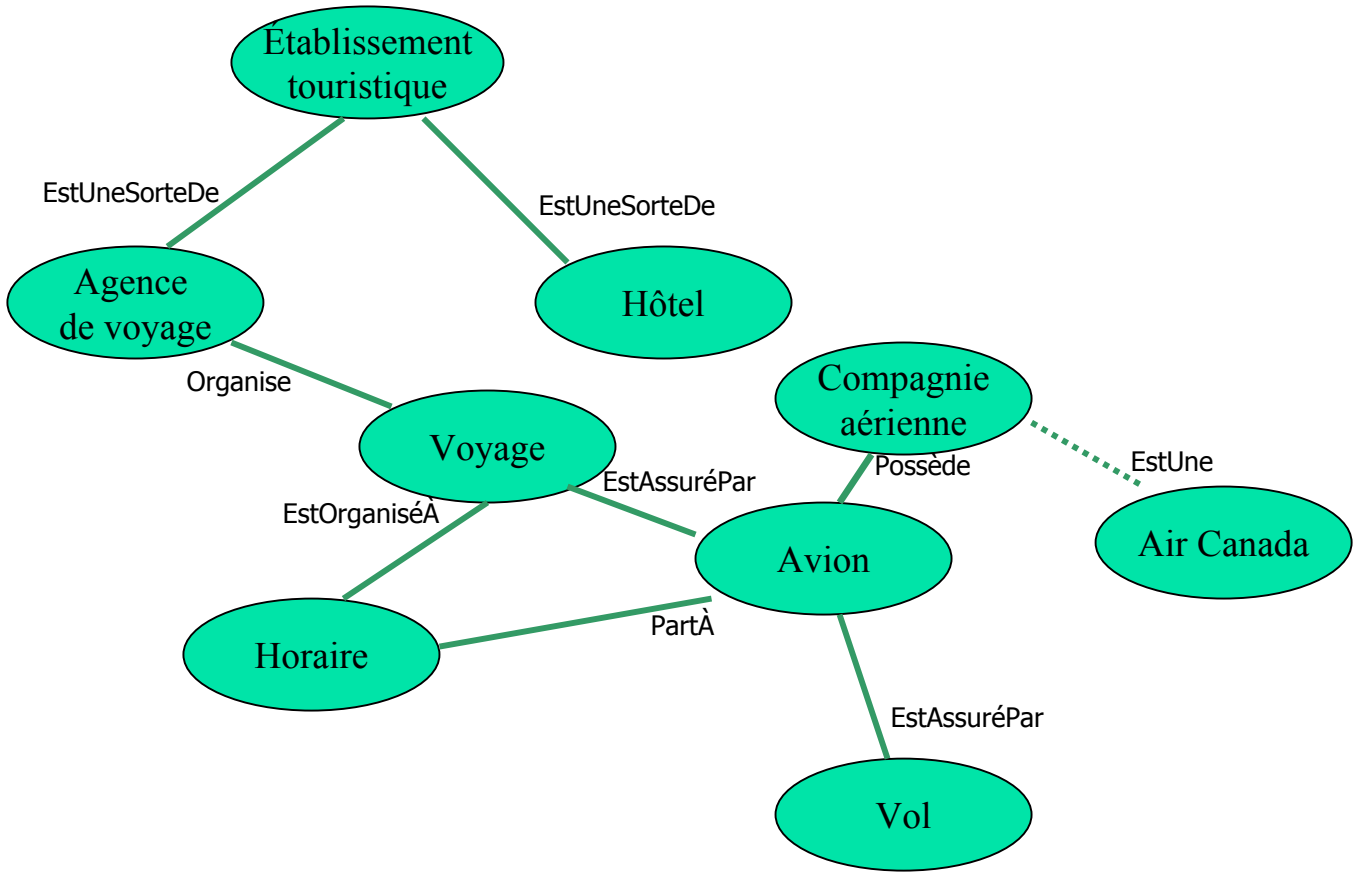
1. Mise en contexte

- Imperfection des moteurs de recherche
- L'annotation de documents se présente comme une solution incontournable
- Ontologie :
  - **Ingénierie des Connaissances** ⟹ une compréhension commune et partagée d'un domaine qui peut être communiquée entre des personnes et des systèmes (Guarino, 1995).
  - **Représentation des connaissances** ⟹ est une spécification formelle et explicite d'une conceptualisation partagée (Gruber, 1993)
- Connaissances intégrées dans les ontologies :

  classes, relations, fonctions, axiomes et instances

Introduction

1. Mise en contexte



Introduction

2. Problématique de recherche

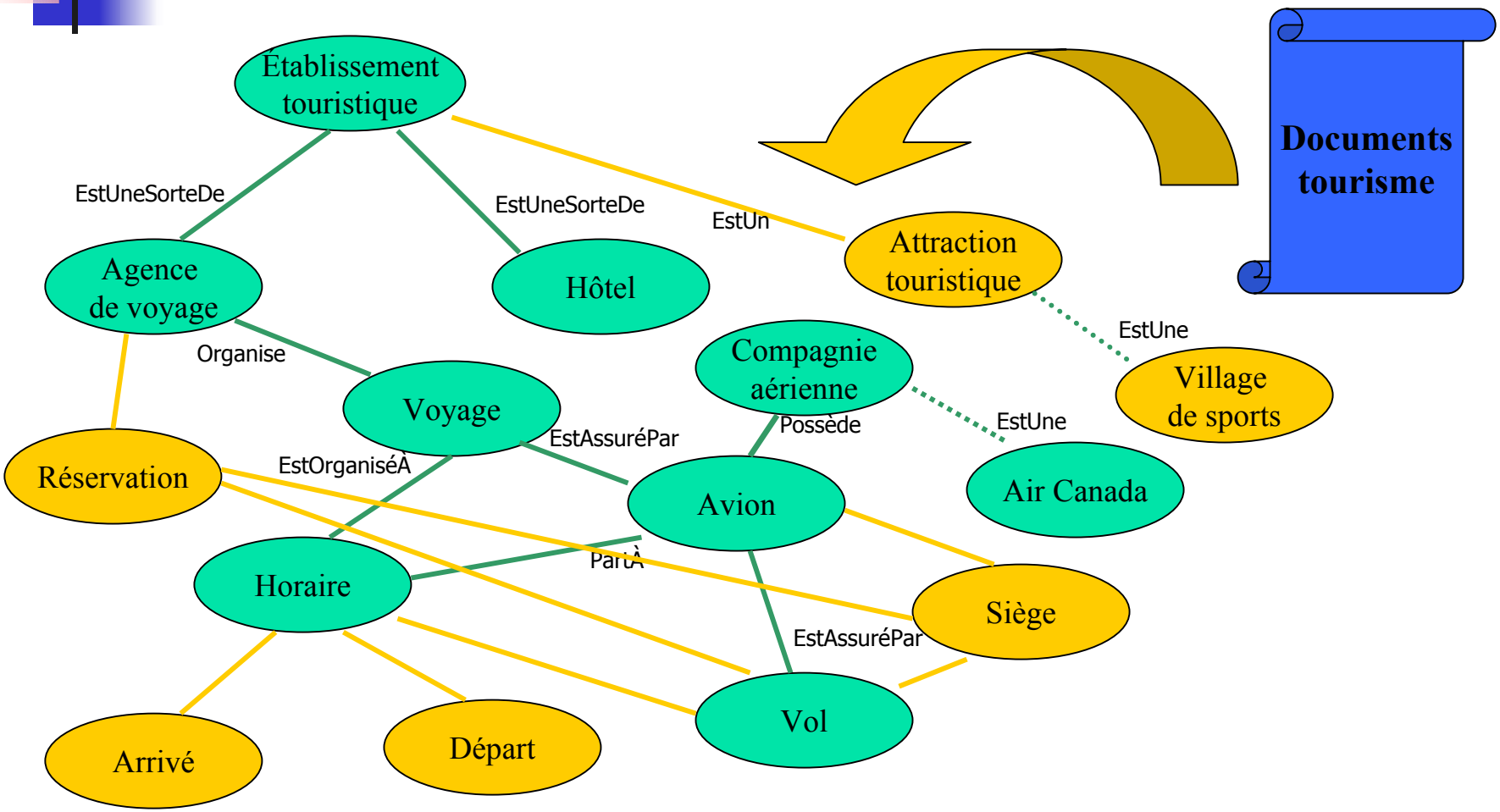
- Problème d'évolution des ontologies

  incomplétudes, erreurs, une nouvelle modélisation du domaine est préférée, le domaine a changé

- Ontologies difficilement réutilisables et partageables
- La majorité des recherches réalisées dans le domaine de l'ingénierie des ontologies se sont concentrées sur les problèmes de construction
- Absence de méthodes consensuelles
- Besoin d'un processus, du moins semi-automatique, de maintenance.

Introduction

1. Mise en contexte



Introduction

3. Objectifs généraux du projet

- Objectif général :

  Mettre en place une « passerelle » entre les documents (lexiques, réseaux sémantiques…) et l'ontologie courante.

- Objectifs spécifiques :
  - Identifier de nouveaux termes clés, spécifiques au domaine
  - Découvrir des relations conceptuelles entre les termes
  - Préciser la place d'un nouveau terme dans l'ontologie
  - Utiliser d'autres sources de données terminologiques
  - Assurer une certaine continuité entre le processus de construction d'ontologie et celui de sa maintenance
  - Assister le cogniticien, chargé de la maintenance, dans la génération de différents états plausibles de changements
  - Proposer un modèle indépendant du domaine
  - Garantir une performance raisonnable en terme de temps d'exécution du système assistant à la maintenance.

Plan de présentation

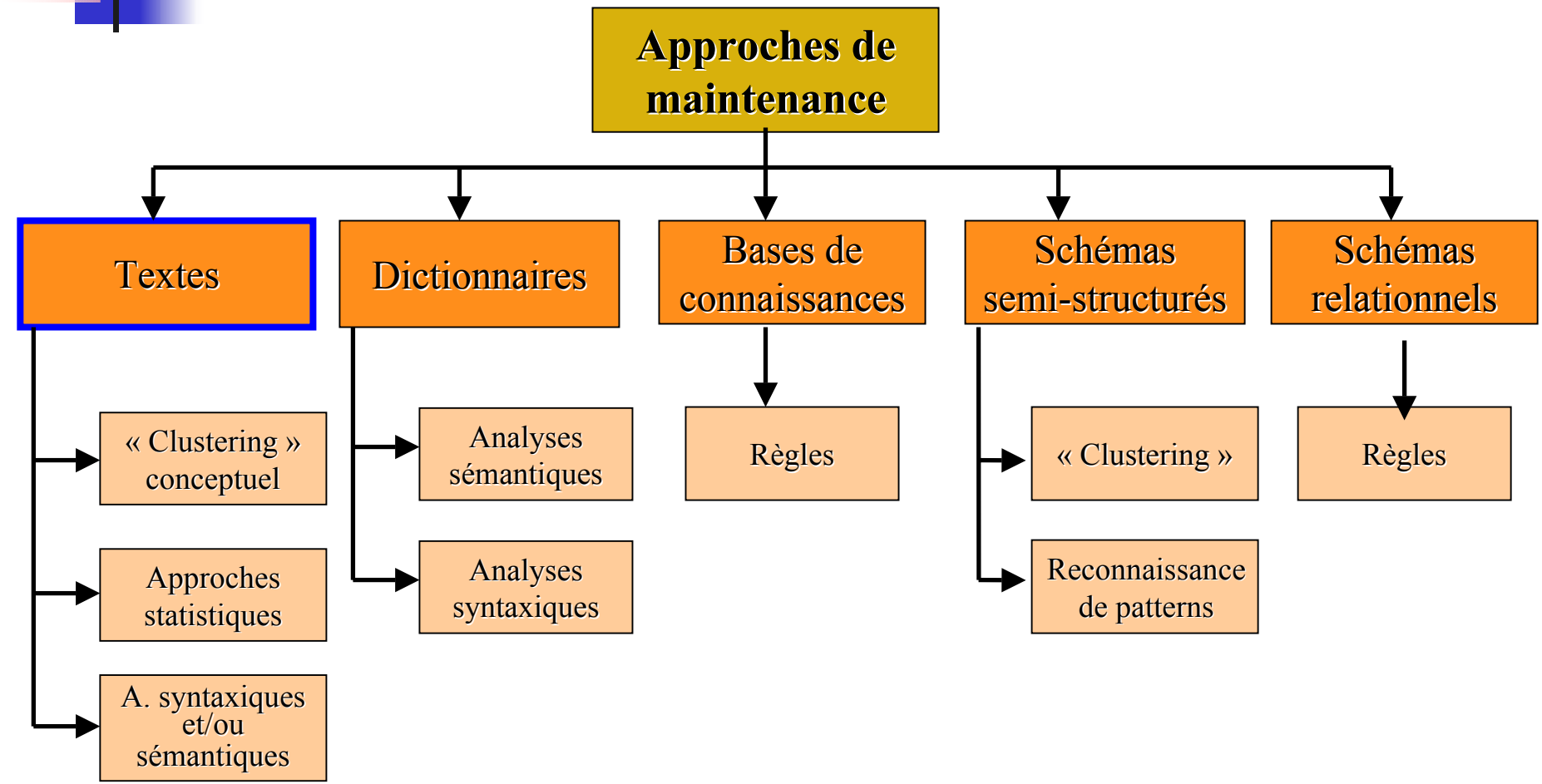
- **Introduction**
- **Problématique**
  - ☞ ***État de l'art***
    - Composante cognitive
    - Composante informatique
    - Choix théoriques
- **Proposition de solution et méthodologie**
  - Modèle proposé
  - Méthodologie de recherche
  - Méthode de validation des résultats
  - Plan sommaire de la thèse
  - État d'avancement des travaux
- **Conclusions**
  - Contributions originales
  - Obstacles à franchir

Problématique

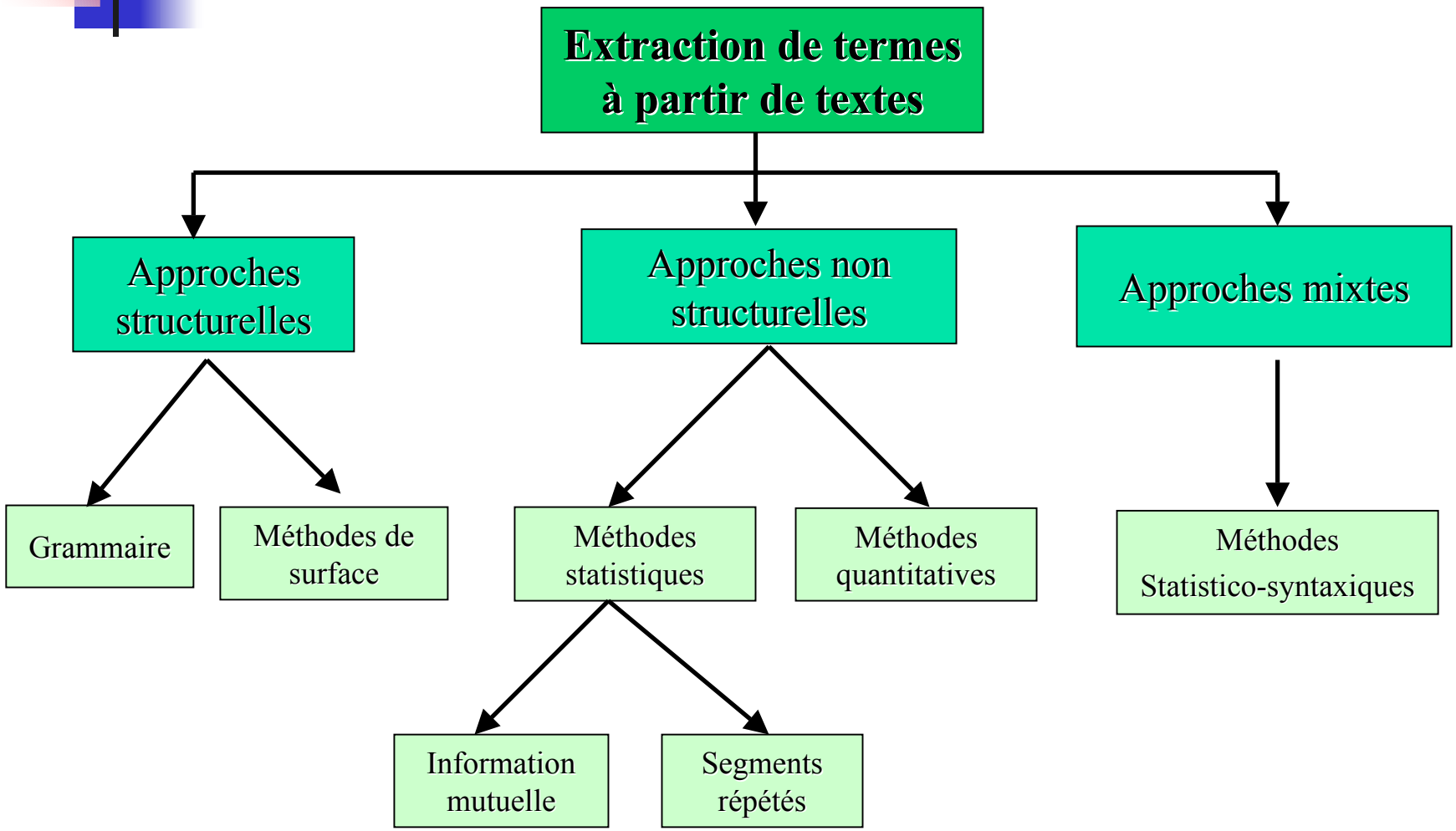
1. État de l'art

- Deux sous-problèmes fondamentaux de la maintenance :
  - Découvrir des termes spécifiques au domaine et pertinents à l'ontologie existante
  - Identifier des relations entre termes
- Variété de techniques :
  - traitement du langage naturel, prospection de données, apprentissage machine et représentation de connaissances.
- La maintenance des ontologies implique un processus d'*apprentissage d'ontologie* :
  - Les recherches n'ont toujours pas atteint un stade de maturité satisfaisant

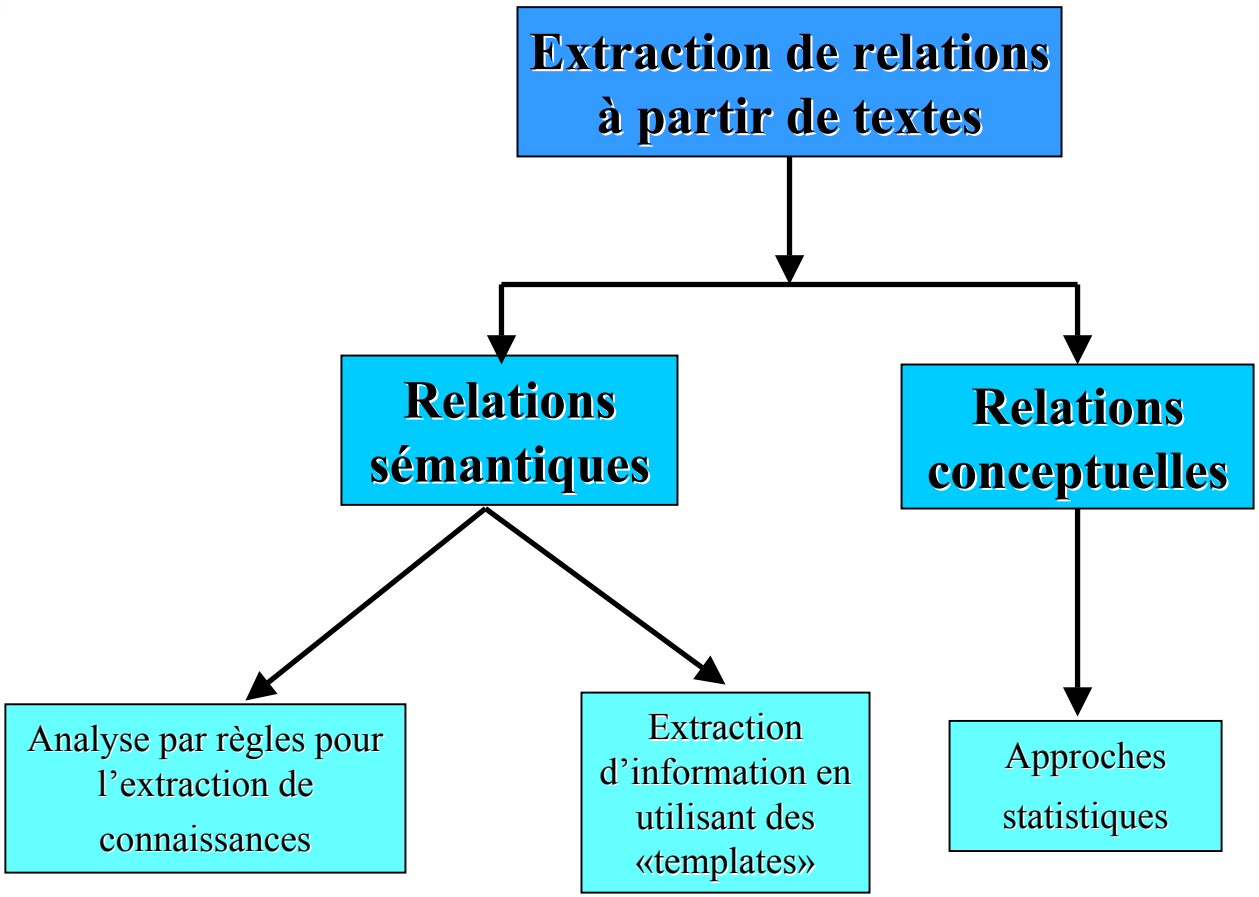
Approches de maintenance



Extraction de termes à partir de textes



Extraction de relations à partir de textes



Plan de présentation

- **Introduction**
- **Problématique**
  - État de l'art

    ☞ ***Composante cognitive***
  - Composante informatique
  - Choix théoriques
- **Proposition de solution et méthodologie**
  - Modèle proposé
  - Méthodologie de recherche
  - Méthode de validation des résultats
  - Plan sommaire de la thèse
  - État d'avancement des travaux
- **Conclusions**
  - Contributions originales
  - Obstacles à franchir

Problématique

2. Composante cognitive

- Certaines des difficultés du TALN ne sont pas dues à des causes contingentes comme la taille de la mémoire, la puissance des microprocesseurs ou la performance des algorithmes, mais plutôt à des conceptions théoriques sur le traitement de données textuelles.
- L'élaboration de systèmes d'assistance à la maintenance des ontologies nécessite l'intégration de processus de raisonnement et d'apprentissage, principalement dédiés à la découverte de relations conceptuelles entre termes à partir de données textuelles.
- L'approche à adopter pour la maintenance des ontologies devrait s'inspirer de façon étroite des mécanismes intellectuels de compréhension, de production et d'apprentissage chez l'être humain (***psycholinguistique***).

Problématique

2. Composante cognitive

- Les recherches en « ***psychologie cognitive*** » montrent que la plupart des mots sont assimilés par la lecture (Landauer et S.T. Dumais, 1997).
  - utilisation des occurrences conjointes d'un terme
  - Exemple : « *micro-processus* » avec : « *électronique* », « *matériel* », « *Unité Centrale de Traitement* », etc
- Application de la technique de classification de documents pour identifier des groupes de termes sémantiquement reliés.
- Notre orientation vers une ***architecture cognitive*** est motivée par un objectif de mise en place d'un système intelligent, supportant des potentialités de l'humain.
  - architecture basée sur différentes sources de connaissances, à savoir :

    les textes, les requêtes des utilisateurs, un thésaurus, et l'ontologie courante.

Plan de présentation

- **Introduction**
- **Problématique**
  - État de l'art
  - Composante cognitive
  - ☞ ***Composante informatique***
  - Choix théoriques
- **Proposition de solution et méthodologie**
  - Modèle proposé
  - Méthodologie de recherche
  - Méthode de validation des résultats
  - Plan sommaire de la thèse
  - État d'avancement des travaux
- **Conclusions**
  - Contributions originales
  - Obstacles à franchir

Problématique

3. Composante informatique

- Réalisation informatique mettant en application le modèle proposé :
  - un « *prototype d'atelier* » d'ingénierie des connaissances, regroupant les modules, déjà existants dans la plate-forme ***SATIM*** ainsi que d'autres modules à développer, et permettant une mise en œuvre efficace de la méthodologie
  - chaîne de traitement sur les textes : « ***ONTOLOGICO*** »
- Les recherches rentrent dans le cadre du projet GDST (*Gestion et Diffusion de Savoir en Télécommunication*)
- La problématique de la maintenance des ontologies constitue une partie cruciale du projet
  - cadre pratique très utile à la validation des résultats de recherches

Plan de présentation

- **Introduction**
- **Problématique**
  - État de l'art
  - Composante cognitive
  - Composante informatique

  ☞ ***Choix théoriques***
- **Proposition de solution et méthodologie**
  - Modèle proposé
  - Méthodologie de recherche
  - Méthode de validation des résultats
  - Plan sommaire de la thèse
  - État d'avancement des travaux
- **Conclusions**
  - Contributions originales
  - Obstacles à franchir

Problématique

4. Choix théoriques du modèle

- **(1) : Application de l'ISL sur des classes de termes**
- **(2) : Ontologie monolingue**
- **(3) : L'ontologie est une conceptualisation partagée**
- **(4) : Indépendance du modèle par rapport à un domaine particulier**
- **(5) : Indépendance du modèle par rapport à la langue**
- **(6) : Importance de la richesse des sources de connaissances**
- **(7) : Cooccurrence : un critère de choix pour le repérage de relations entre termes**
- **(8) : La cooccurrence résout, dans une certaine mesure, le problème d'ambiguïté lexicale**
- **(9) : Les significations de termes contribuent activement au repérage de relations entre termes**
- **(10) : L'intervention d'un expert est une opération incontournable**
- **(11) : Importance de la complétude des hypothèses de relations entre termes**

Plan de présentation

- **Introduction**
- **Problématique**
  - État de l'art
  - Composante cognitive
  - Composante informatique
  - Choix théoriques
- **Proposition de solution et méthodologie**
  - ☞ ***Modèle proposé***
  - Méthodologie de recherche
  - Méthode de validation des résultats
  - Plan sommaire de la thèse
  - État d'avancement des travaux
- **Conclusions**
  - Contributions originales
  - Obstacles à franchir

Proposition de solution et méthodologie

1. Modèle proposé

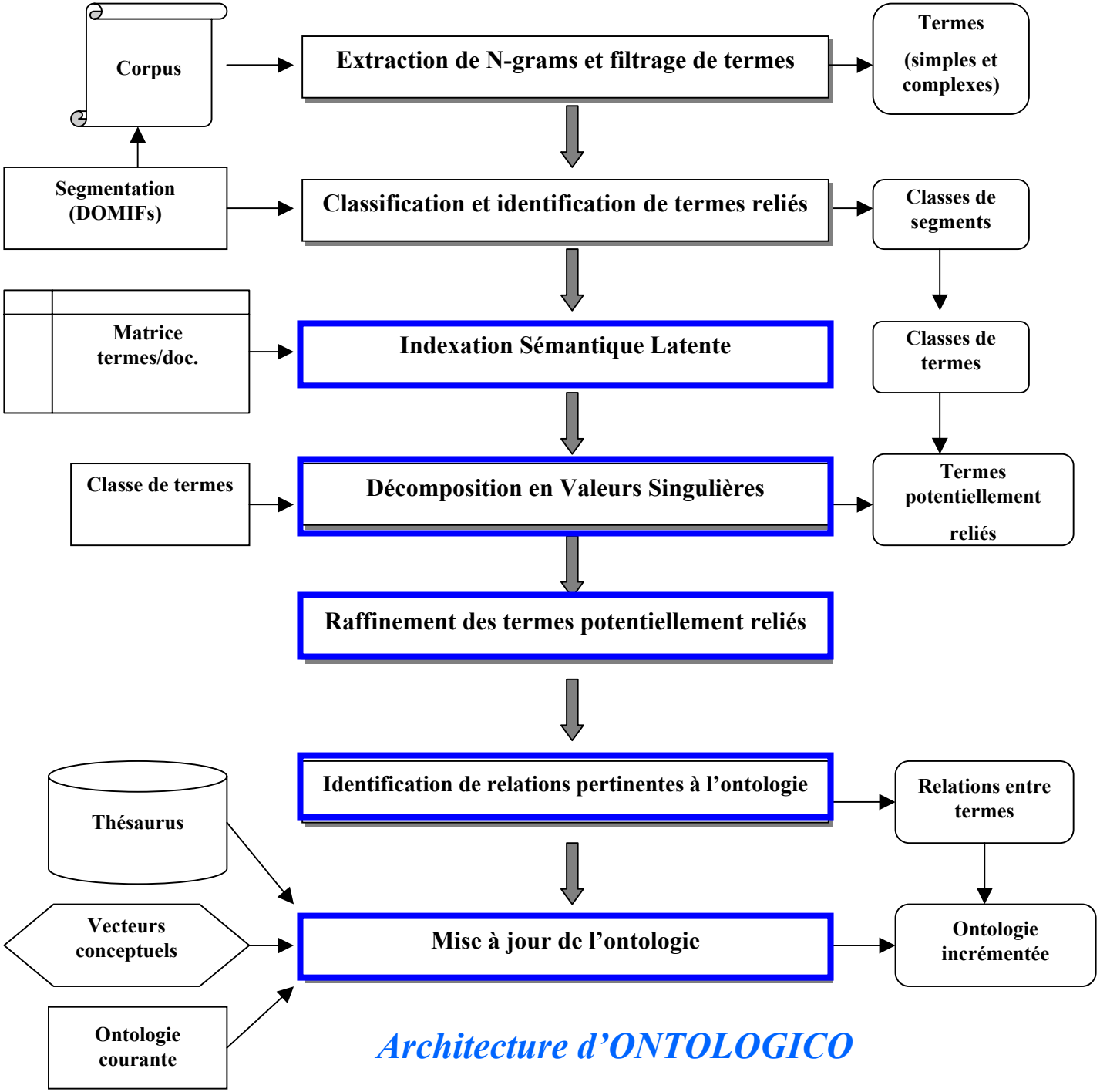
- **Objectif** : fournir de l'assistance à l'utilisateur pour assurer un raffinement continuel de l'ontologie tout en assurant la consistance et la cohérence de celle-ci et de ses artefacts
- « ***Raffinement Conceptuel par Analyse Vectorielle*** » (RCAV)
  - données textuelles
  - classification de textes
  - Indexation Sémantique Latente (ISL) (Deerwester and al. 1990)
  - ressource terminologique (thésaurus)
  - analyse vectorielle
  - intervention humaine

Proposition de solution et méthodologie

1 Modèle proposé

Processus itératif d'ingénierie :

- 1- Extraction de N-grams et filtrage de termes
- 2- Classification et identification de termes reliés
- 3- Indexation Sémantique Latente
- 4- Décomposition en Valeurs Singulières
- 5- Raffinement des termes potentiellement reliés
- 6- Identification de relations entre termes pertinentes à l'ontologie
- 7- Mise à jour de l'ontologie



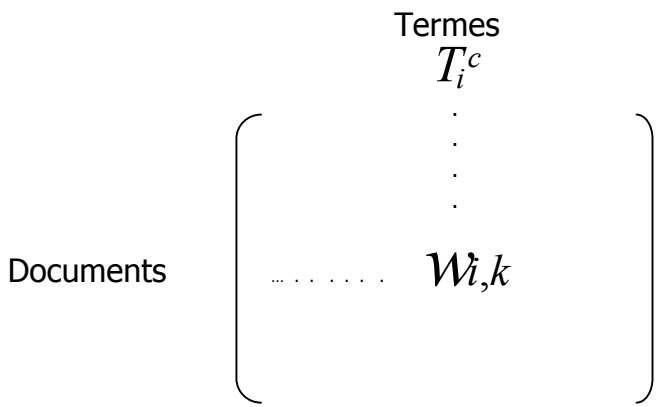
Architecture d'ONTOLOGICO

Proposition de solution et méthodologie

1. Modèle proposé

Indexation Sémantique Latente

- Objectif : extraire, à partir de ces classes de termes identifiées, ceux représentant un niveau élevé de corrélation
- L'ISL représente les documents par des concepts qui sont réellement et statistiquement indépendants de telle sorte que les termes ne le sont pas
- ISL (Deerwester and al. 1990, Srivastava and al. 2002) utilisée, spécialement pour sa simplicité et sa justification par des fondements mathématiques précis.
- Entrée : matrice termes-documents.



Proposition de solution et méthodologie

1. Modèle proposé

- Poids des termes : $$w_{i,k} = \frac{C_{i,k}}{\sum_{j=1}^{n_k} C_{j,k}}$$
- Normalisation des poids (Greengrass, 1997) : $$W_{i,k} = \frac{w_{i,k}}{\sqrt{\sum_{j=1}^{n_k} w_{j,k}^2}}$$

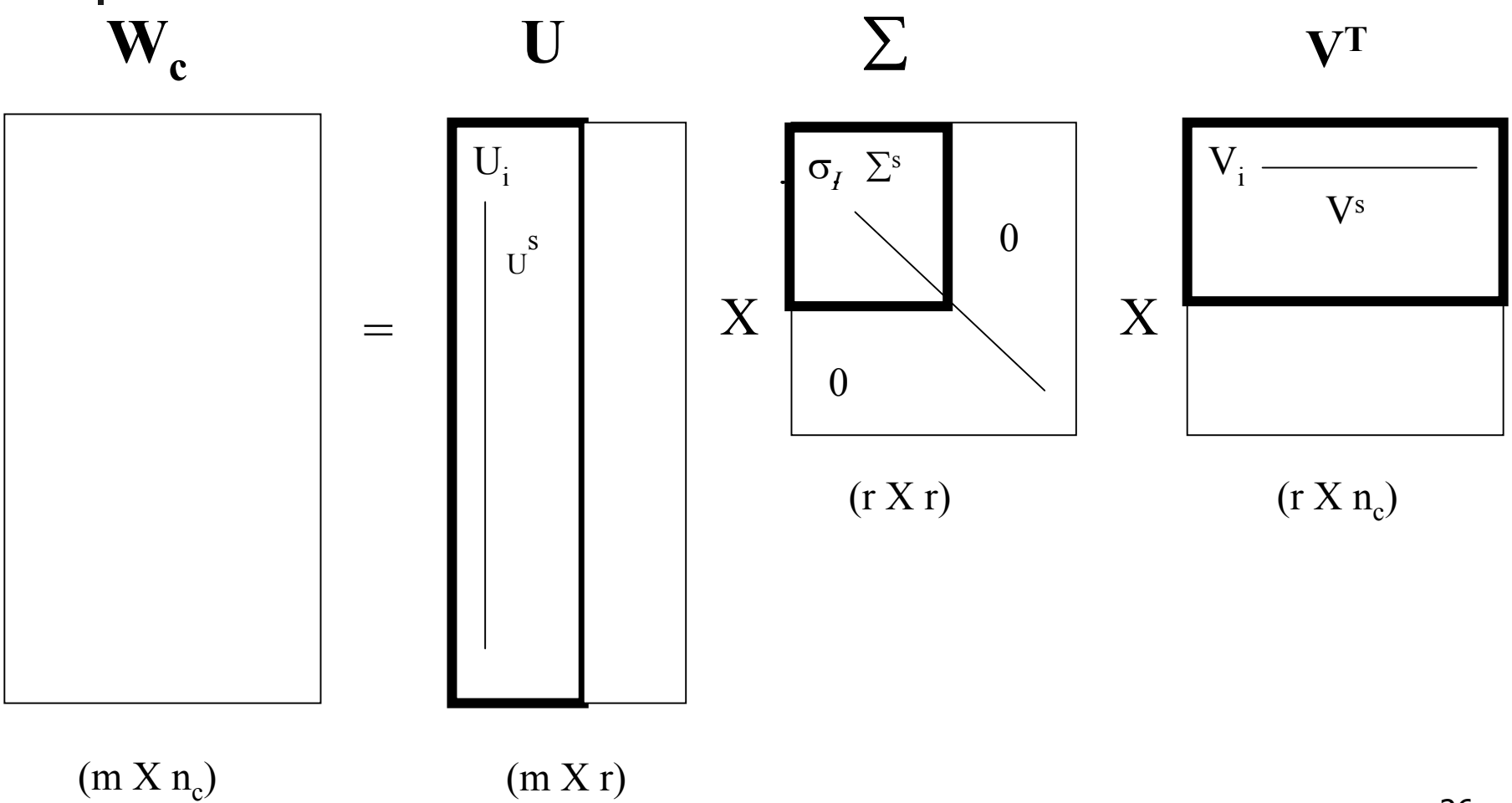
Décomposition en Valeurs Singulières

- DVS : (Golub et al., 1969)
- Transformation de formule n'entraînant aucune perte de généralité :

$$W_c = U \Sigma V^T$$

Proposition de solution et méthodologie

1. Modèle proposé



Proposition de solution et méthodologie

1. Modèle proposé

- Éliminer toutes les valeurs singulières de Σ inférieures à un seuil de pourcentage de la valeur singulière la plus large, σ_1

  W_c^S : approximation de W_c :

$$W_c^s = U^s \Sigma^s V^{sT}$$

- U^s semble être le composant le plus important pour nous.

  ⇨ la matrice (m X s) représente les corrélations entre les termes dans la collection de documents et appartenant à la classe *c*.

- L'application de l'ISL sur les groupes de termes identifiés par la classification, plutôt que sur la totalité des termes du texte, possède l'avantage de réduire la matrice de cooccurrences de termes dans les documents à une dimension raisonnable.

Proposition de solution et méthodologie

1. Modèle proposé

Identification de relations entre termes pertinentes à l'ontologie

- Méthode fondée sur un mariage entre, d'une part, l'approche *des réseaux sémantiques* (exp. : *WordNet*) associée au domaine de la « représentation des connaissances », et d'autre part, *l'approche vectorielle* issue des « représentations saltoniennes » (Salton, 1968) et de la recherche d'information.
- Les thésauri sont spécialement utiles pour offrir des réseaux lexicaux et de l'information additionnelle reliée à la signification de termes (utilisation, définition, synonymie, etc.).
- Termes représentés par des vecteurs conceptuels (construits à partir des items lexicaux associés à chacun de ces termes).
- Mesures de proximité sémantique entre vecteurs conceptuels

  - Exemple : la « *distance thématique* »

Proposition de solution et méthodologie

1. Modèle proposé

Mise à jour de l'ontologie

- Termes représentés par des vecteurs conceptuels (construits à partir des items lexicaux associés à chacun de ces termes).
- Les nouveaux termes, ne figurant pas dans l'ontologie courante, ainsi que leurs relations, sont intégrés à cette ontologie.
- Un ensemble de règles assurant la cohérence du modèle global doit être respecté.
- Étiquetage des relations par un expert.
- Reprise des itérations pour chacune des classes de termes identifiées.

Plan de présentation

- **Introduction**
- **Problématique**
  - État de l'art
  - Composante cognitive
  - Composante informatique
  - Choix théoriques
- **Proposition de solution et méthodologie**
  - Modèle proposé
  - ☞ ***Méthodologie de recherche***
  - Méthode de validation des résultats
  - Plan sommaire de la thèse
  - État d'avancement des travaux
- **Conclusions**
  - Contributions originales
  - Obstacles à franchir

Proposition de solution et méthodologie

2. Méthodologie de recherche

Justification du modèle proposé

- Exploration des approches et des outils qui se rattachent d'une façon directe ou indirecte au problème posé.
- L'analyse critique des approches existantes, de leurs points forts et points faibles nous a permis de formuler un ensemble de choix théoriques.
- La capacité des techniques statistiques, à traiter de larges données textuelles.
- La technique de l'ISL a spécialement montré sa fiabilité.
- Associer l'ISL à la classification textuelle.
- Les données textuelles ne peuvent, toutes seules, supporter la modélisation d'un domaine ⇨ utilisation d'un thésaurus .
- Représentation de termes par des vecteurs conceptuels et utilisation de mesures de similarité.

Proposition de solution et méthodologie

2. Méthodologie de recherche

Constitution de corpus

- Sélection à partir d'une documentation technique relative au domaine ; télécommunications sans fils
- Historique des requêtes formulées par les utilisateurs.

Expérimentation et développement

- Le modèle proposé sera implémenté sous forme d'une chaîne de traitements (**ONTOLOGICO**) au sein de la plate-forme SATIM.

  Expérimentation de différents types d'analyses grâce à sa modularité, sa flexibilité, ses diverses fonctions d'analyse et sa capacité d'adaptation par rapport à la croissance des données textuelles

- Évaluation de l'assistance fournie par notre système aux utilisateurs.

  Comparer la version incrémentée de l'ontologie à une ontologie de référence.

- Développement en langage C++.
- L'aspect modulaire de l'outil est un choix pratique particulièrement puissant.

Plan de présentation

- **Introduction**
- **Problématique**
  - État de l'art
  - Composante cognitive
  - Composante informatique
  - Choix théoriques
- **Proposition de solution et méthodologie**
  - Modèle proposé
  - Méthodologie de recherche
  - ☞ ***Méthode de validation des résultats***
  - Plan sommaire de la thèse
  - État d'avancement des travaux
- **Conclusions**
  - Contributions originales
  - Obstacles à franchir

Proposition de solution et méthodologie

3. Méthode de validation des résultats

- Évaluer une technique d'apprentissage d'ontologie revient à mesurer la similarité entre une ontologie, manuellement conçue, considérée en tant que « *standard de référence* » et une ontologie générée en utilisant cette technique.
- Processus d'évaluation de la maintenance souvent manuel

  (Bachimont et al., 2002), (Faatz and Steinmetz, 2002), (Gupta et al., 2002), (Hearst, 1998), (Hwang, 1999), (Khan and Luo, 2002), (Kietz et al., 2000), (Missikoff et al., 2002), (Moldovan and Girju, 2001), etc.
- Évaluation en trois étapes
  - Apprentissage de l'ontologie dans sa totalité
  - Le niveau lexical
  - Le niveau conceptuel

Proposition de solution et méthodologie

3. Méthode de validation des résultats

- **Précision :**

  mesure de la proportion des éléments corrects sélectionnés par le système

$$\text{Précision} = \frac{Comp \text{ } \text{Réf}}{Comp}$$

- **Rappel :**

  mesure standard de la quantité d'éléments repérés

$$\text{Rappel} = \frac{Comp \text{ } \text{Réf}}{\text{Réf}}$$

- **Niveau lexical :** (Maedche A, Staab S., 2001)

  SM : « *mesure de similarité lexicale* » entre l_i et l_j.
  ed : « *la distance d'édition* » (Levenshtein, 1996)

  $\overline{SM}$: «*l'appariement entre chaînes de caractères*»: lexiques L_1 (Réf) et L_2 des 2 ontologies

$$SM(l_i, l_j) = \max\left(0, \frac{\min(|l_i|, |l_j|) - ed(l_i, l_j)}{\min(|l_i|, |l_j|)}\right) \in [0, 1]$$

$$\overline{SM}(L_1, L_2) = \frac{1}{|L_1|} \sum_{l_i \in L_1} \max_{l_j \in L_2} SM(l_i, l_j)$$

Proposition de solution et méthodologie

3. Méthode de validation des résultats

- **Niveau conceptuel :** (Maedche A, Staab S., 2001)

comparaison de structures sémantiques d'ontologies O_1 et O_2

$\overline{TO}$: *« mesure moyenne de similarité » entre 2 taxonomies* H_1^c *et* H_2^c

$SC(C_i, H^c)$: *mesure «**Conceptual Cotopy** » du terme Ci : l'ensemble de tous les termes aux niveaux supérieur et inférieur à Ci*

$$SC(C_i, H^c) = \{ C_j \in C \mid H^c(C_i, C_j) \text{ ✞ } H^c(C_j, C_i) \text{ ✞ } C_i = C_j \}$$

Exemple : la mesure SC du terme « personne » est donnée par :

$$F_1^{-1}(SC(F(\{"personne"\}), H^c)) = \{"étudiant", "chercheur", "personne"\}$$

Proposition de solution et méthodologie

3. Méthode de validation des résultats

$$\overline{TO}(O_1,O_2)=\frac{1}{|L_1^c|}\sum_{L\in L_1^c} TO(L,O_1,O_2)$$

$$TO(L,O_1,O_2)=\begin{cases} TO'(L,O_1,O_2) & \text{Si} \quad L\in L_2^c \\ TO''(L,O_1,O_2) & \text{Si} \quad L\notin L_2^c \end{cases}$$

$$TO'(L',O_1,O_2)=\frac{|F_1^{-1}(SC(F(\{L'\}),H_1^c))\cap F_2^{-1}(SC(F(\{L'\}),H_2^c))|}{|F_1^{-1}(SC(F(\{L'\}),H_1^c))\cup F_2^{-1}(SC(F(\{L'\}),H_2^c))|}$$

$$TO''(L'',O_1,O_2)=\max_{c\in c2}\frac{|F_1^{-1}(SC(F(\{L''\}),H_1^c))\cap F_2^{-1}(SC(C),H_2^c)|}{|F_1^{-1}(SC(F(\{L''\}),H_1^c))\cup F_2^{-1}(SC(C),H_2^c)|}$$

Plan de présentation

- **Introduction**
- **Problématique**
  - État de l'art
  - Composante cognitive
  - Composante informatique
  - Choix théoriques
- **Proposition de solution et méthodologie**
  - Modèle proposé
  - Méthodologie de recherche
  - Méthode de validation des résultats

  ☞ ***Plan sommaire de la thèse***
  - État d'avancement des travaux
- **Conclusions**
  - Contributions originales
  - Obstacles à franchir

Proposition de solution et méthodologie

4. Plan sommaire de la thèse

- **Chapitre 1 : Introduction**
  - description générale de la problématique de la maintenance des ontologies
  - les objectifs généraux de notre projet de recherche.
  - les champs de recherche
- **Chapitre 2 : Problématique**
  - état de l'art de la problématique, des approches d'extraction de termes à partir de textes et des techniques de repérage de relations entre termes.
  - la composante cognitive (l'analyse sémantique, la psychologie cognitive et les fondements d'une architecture cognitive)
  - hypothèses du modèle

Proposition de solution et méthodologie

4. Plan sommaire de la thèse

- **Chapitre 3 : Solution proposée et méthodologie**
  - Nos options théoriques : la psychologie cognitive, la cooccurrence, la classification textuelle, l'Indexation Sémantique Latente et les vecteurs conceptuels
  - modèle proposé
  - outils utilisés
  - méthodologie
- **Chapitre 4 : Implémentation et évaluation**
  - spécifications relatives à la chaîne de traitement **ONTOLOGICO**.
  - validation de notre approche
- **Chapitre 5 : Conclusions et perspectives**
  - synthèse du travail réalisé
  - les contributions originales du projet
  - les éventuelles critiques et limites touchant la solution
  - perspectives dans le domaine de l'ingénierie des connaissances

Plan de présentation

- **Introduction**
- **Problématique**
  - État de l'art
  - Composante cognitive
  - Composante informatique
  - Choix théoriques
- **Proposition de solution et méthodologie**
  - Modèle proposé
  - Méthodologie de recherche
  - Méthode de validation des résultats
  - Plan sommaire de la thèse

  ☞ ***État d'avancement des travaux***
- **Conclusions**
  - Contributions originales
  - Obstacles à franchir

Proposition de solution et méthodologie

5. État d'avancement des travaux

- **État de l'art**
- **Proposition de modèle**

- Gargouri, Y., Lefebvre, B. et Meunier, J.G. (2003a) «***Maintenance des ontologies à partir d'analyses textuelles***» ACFS'2003, Rimouski, Québec, 21 Mai 2003.
- Gargouri, Y., Lefebvre, B. and Meunier, J.G. (2003b) « ***Ontology Maintenance using Textual Analysis*** », SCI'2003 : The 7th World Multiconference Systems, Cybernetics and Informatics, Orlando, Florida, July 2003 – And Systems, Cybernetics and Informatics Journal.
- Gargouri, Y., Lefebvre, B. et Meunier, J.G. (2004) «***ONTOLOGICO : vers un outil d'assistance au développement itératif des ontologies*** ». Journées d'études sur Terminologie, Ontologie, et Représentation des connaissances (TERMINO'2004). Lyon, France, janvier 2004.

Plan de présentation

- **Introduction**
- **Problématique**
  - État de l'art
  - Composante cognitive
  - Composante informatique
  - Choix théoriques
- **Proposition de solution et méthodologie**
  - Modèle proposé
  - Méthodologie de recherche
  - Méthode de validation des résultats
  - Plan sommaire de la thèse
  - État d'avancement des travaux
- **Conclusions**
  - ☞ ***Contributions originales***
  - Obstacles à franchir

Conclusions

1. Contributions originales

- L'application de la technique de ISL sur les groupes de termes identifiés par la classification, plutôt que sur la totalité des termes du texte.
- «***Raffinement Conceptuel par Analyse Vectorielle***» (RCAV): un processus cohérent de raffinement graduel de relations conceptuelles entre termes.
- Une méthodologie et des outils pour maintenir une ontologie à partir d'analyses textuelles : remèdes aux lacunes méthodologiques.
- Contribution aux travaux dans le domaine de l'extraction de connaissances à partir d'analyses statistiques de textes.
- Une réflexion sur les divers paliers à envisager dans une démarche de modélisation de connaissances textuelles pour des objectifs de maintenance d'une ontologie.

Conclusions

1. Contributions originales

- Indépendance de la langue et du domaine.
- Combinaison de diverses sources de connaissances (corpus, thésaurus, requêtes)

Plan de présentation

- **Introduction**
- **Problématique**
  - État de l'art
  - Composante cognitive
  - Composante informatique
  - Choix théoriques
- **Proposition de solution et méthodologie**
  - Modèle proposé
  - Méthodologie de recherche
  - Méthode de validation des résultats
  - Plan sommaire de la thèse
  - État d'avancement des travaux
- **Conclusions**
  - Contributions originales

☞ ***Obstacles à franchir***

Conclusions

2. Obstacles à franchir

- Affronter la complexité du traitement des données textuelles en vue d'exploiter la richesse implicite de cette source de connaissances.
- L'identification complète et exacte de termes appartenant à un domaine spécifique est considérée comme un pré-traitement d'une grande importance pour la production de résultats adéquats et fiables .
- Les données textuelles que nous traitons sont de larges tailles : problème du temps d'exécution.
- Problème d'ambiguïté sémantique. Exp. : les homonymes, etc.
- Problème de subjectivité des experts lors des processus de validation.

Questions?