

Indexation automatique pour le Web sémantique

Fabien Barbas

Département Génie Informatique
École Polytechnique de Montréal

fabien.barbas@polymtl.ca

MAPI – Juin 2006

Plan de la présentation

- Introduction
- Objectifs de recherche
- Présentation du projet
- Conclusion

Plan de la présentation

- Introduction
 - Problématique
 - Contexte
 - Web sémantique, ontologies et indexation
- Objectifs de recherche
- Présentation du projet
- Conclusion

Définition du Web sémantique

"The Semantic Web is an **extension** of the current Web in which information is given **well-defined meaning**, better enabling *computers* and people to work in cooperation."

-- Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, *Scientific American*, Mai 2001.

Technologies du Web sémantique

- Métadonnées
- Ontologies
- Moteurs d'inférence
- Agents

Ontologie

- Qu'est-ce qu'une ontologie?
- À quoi ça sert?

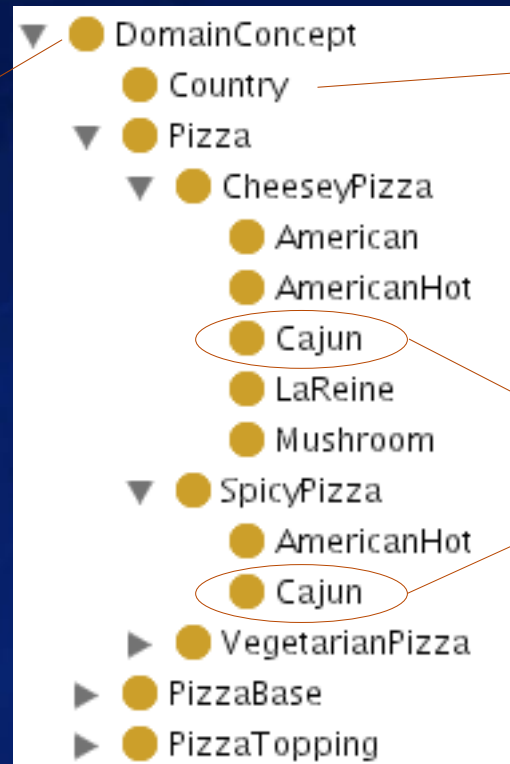
“An ontology is an explicit and formal specification of a conceptualization.”

-- T.R. Gruber

Exemple d'ontologie

Ontologie de pizzas!

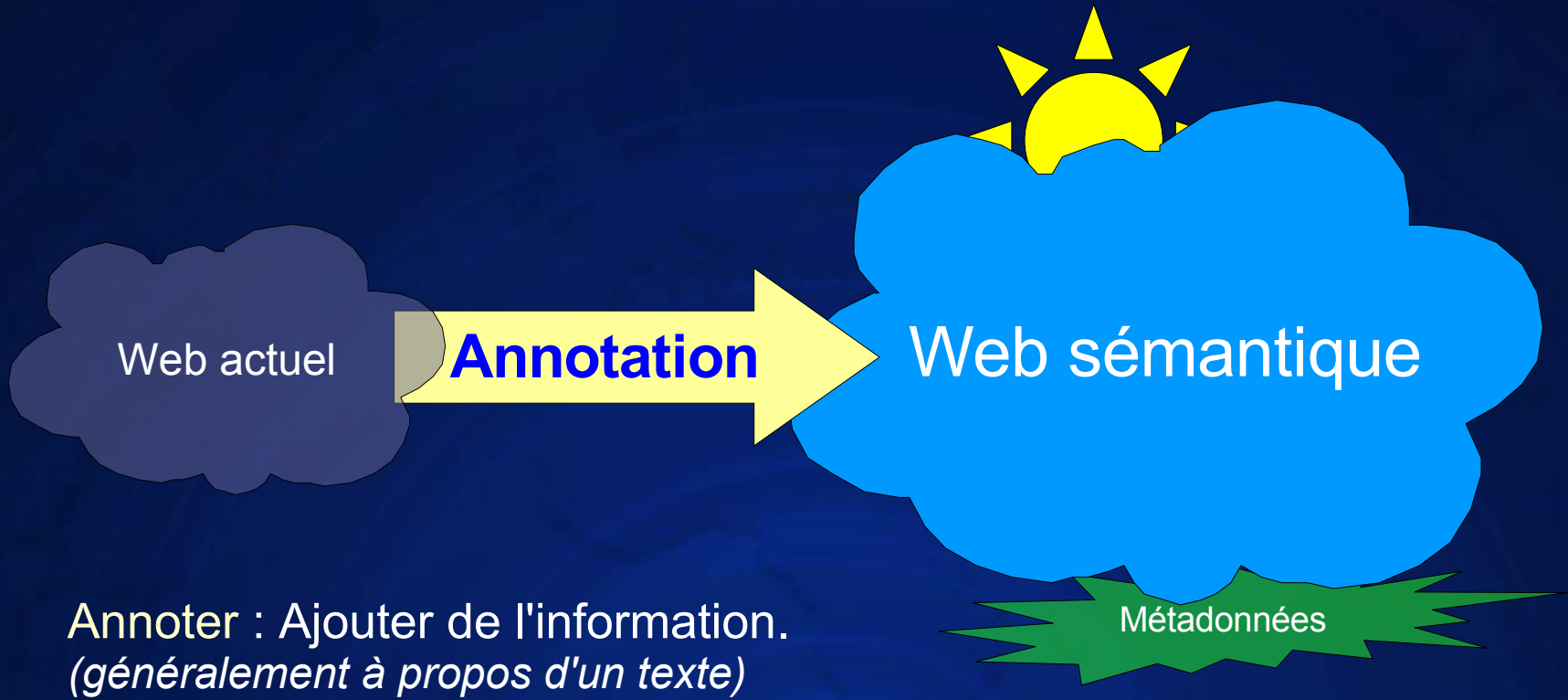
Hiérarchie
de concepts



On peut
représenter
différent
domaines

Graphe

Annotation



Indexation

- Méthode qui permet d'indexer un document
- Identification d'éléments dans un document
 - Indexation libre
 - Détermine les éléments intéressants
 - Indexation contrainte
 - Identifie des éléments d'un vocabulaire prédéfini
- Fait partie du processus d'annotation



Indexation manuelle vs automatique

- Manuelle

- Facile
- Long
- Subjectif (biaisé)
- Sujet au SPAM
- Peu fiable...

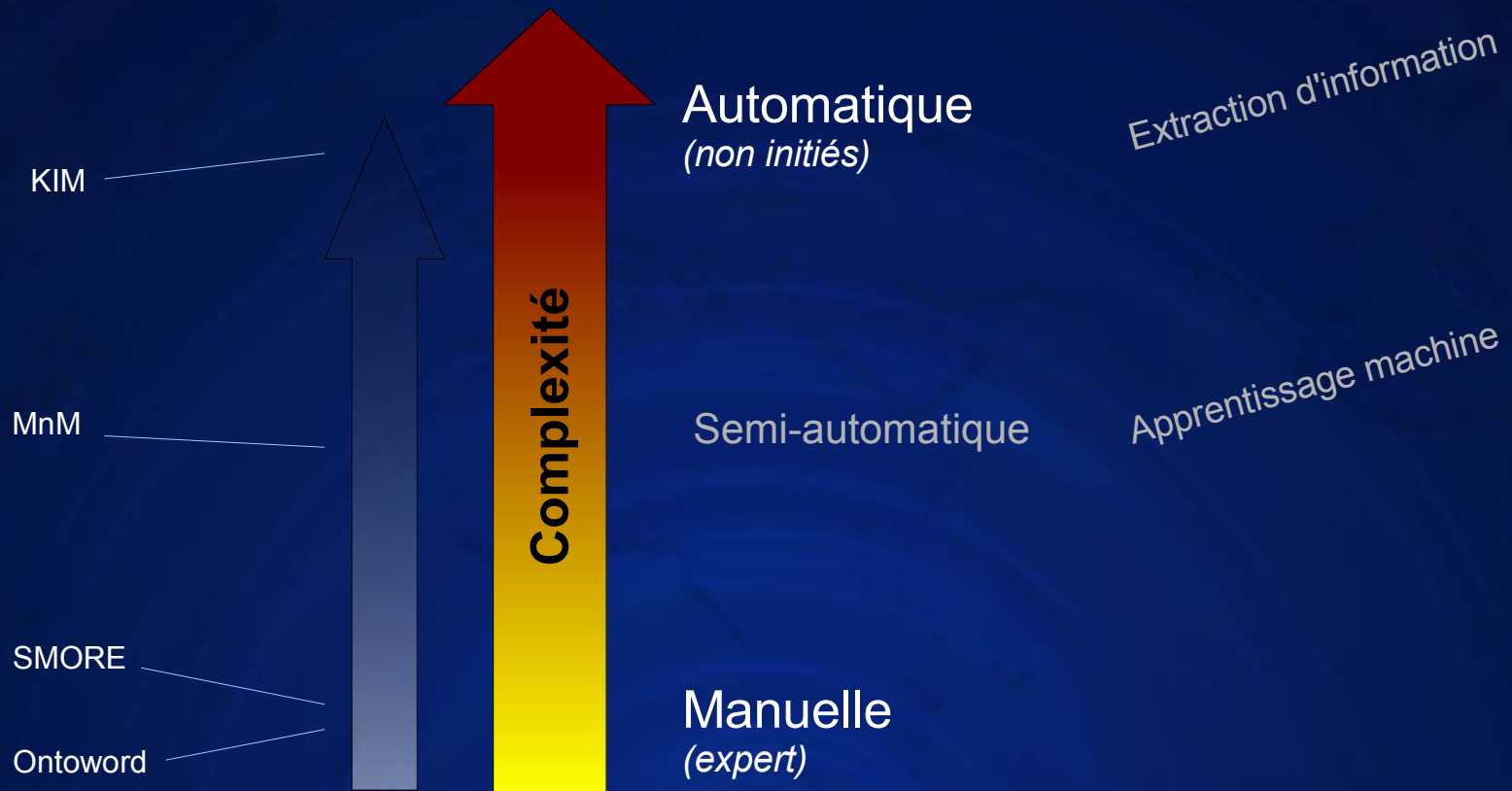
- Automatique

- Difficile
- Rapide
- Objectif *

Difficulté de juger de la pertinence

- Résultats sur le jugement humain
 - Indexer un même document à l'aide d'un thésaurus
 - Répondre à des questions en cherchant une base de données
 - Juger de la pertinence d'un ensemble de documents par rapport à une question
- Indexation manuelle
 - Non initiés : 25-35%
 - Experts : ~15%

Automatisation vs Complexité



Plan de la présentation

- Introduction
- Objectifs de recherche
 - Solution d'indexation automatique
 - NLP
- Présentation du projet
- Conclusion

Objectifs de recherche

- Développer une solution d'indexation
 - Basée sur des concepts (ontologie)
 - Utilisable par des non initiés (simple)
 - Entièrement automatique
 - Générique (indépendante des sources)
 - Doit donner de bons résultats (vs. humain)

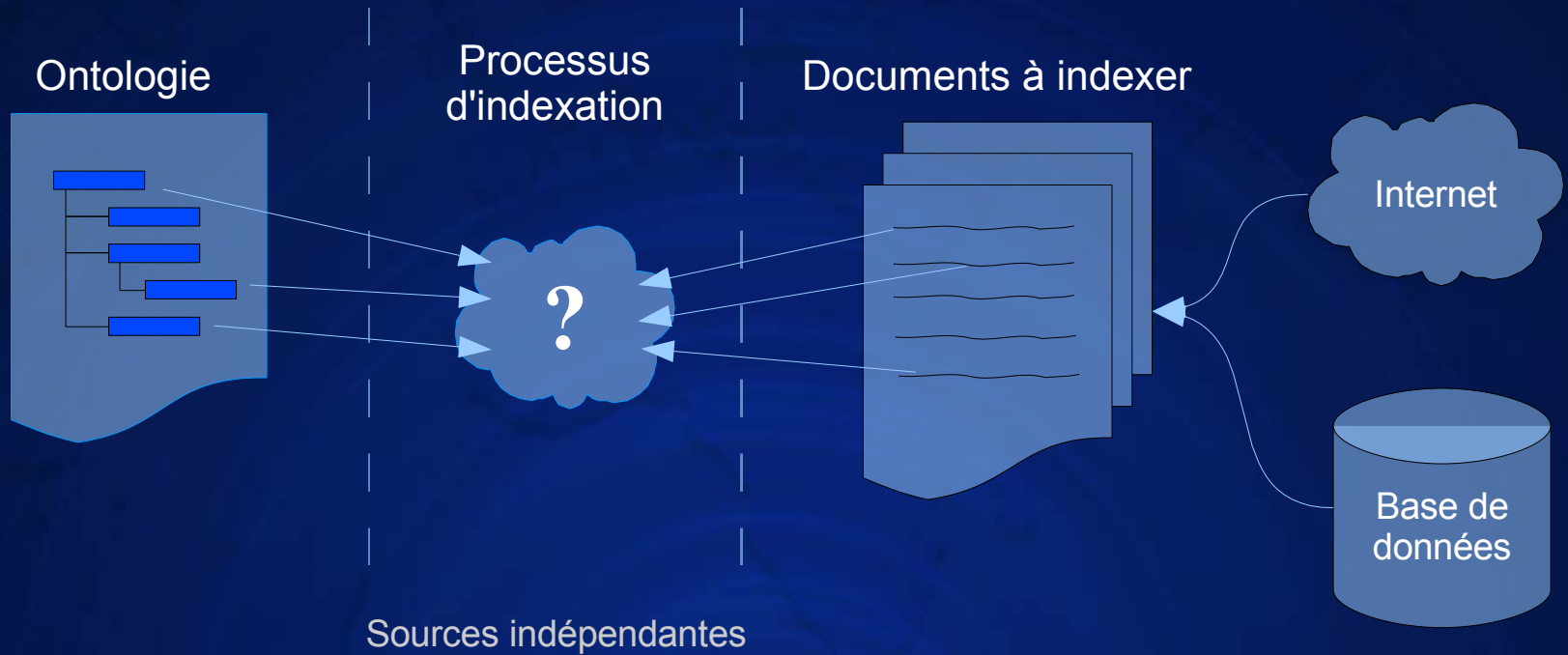
Objectif secondaire

- Explorer le traitement de la langue naturelle (TLN)
 - Évaluation de l'influence des paramètres linguistiques
 - Multilingue (Français, anglais, portugais, etc.)

Plan de la présentation

- Introduction
- Objectifs de recherche
- **Présentation du projet**
 - Projet
 - Résultats
- Conclusion

Schéma général du projet



Tests pratiques

- **Ontologie**

- ACM (*Association for Computing Machinery*)

- Taxonomie des sujets en informatique

- ~1500 concepts étiquetés à la main par des experts du domaine

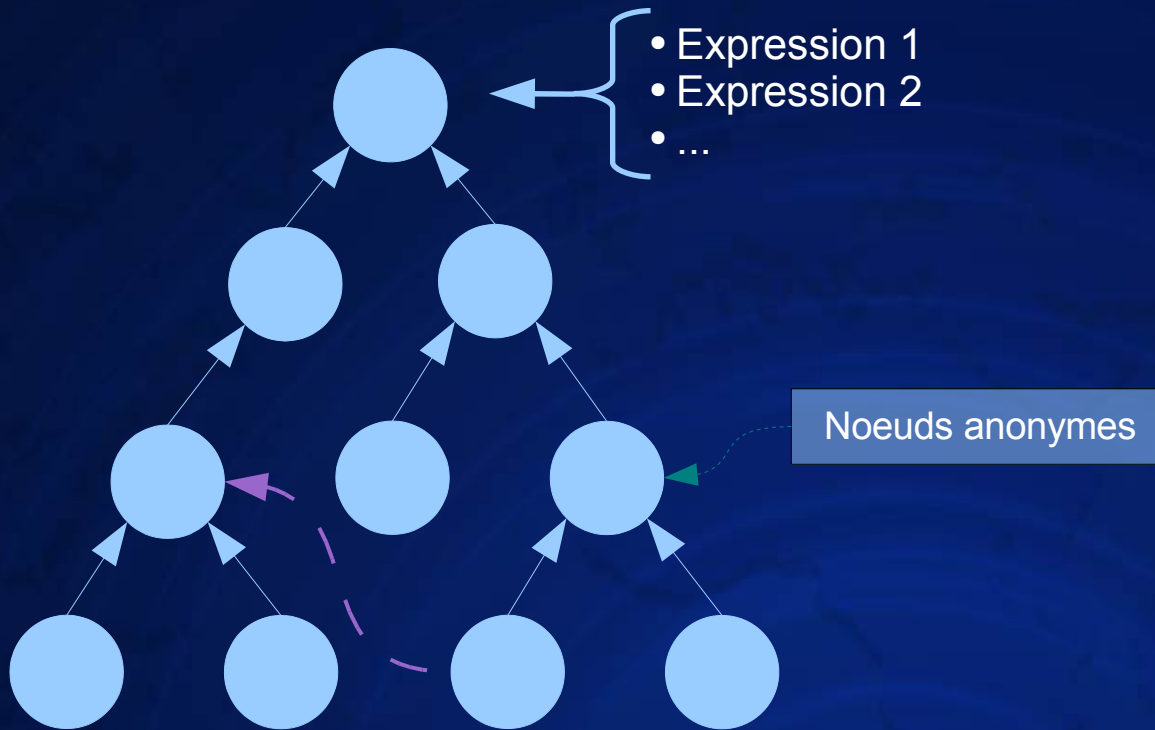
- Département de génie informatique

- **Documents**

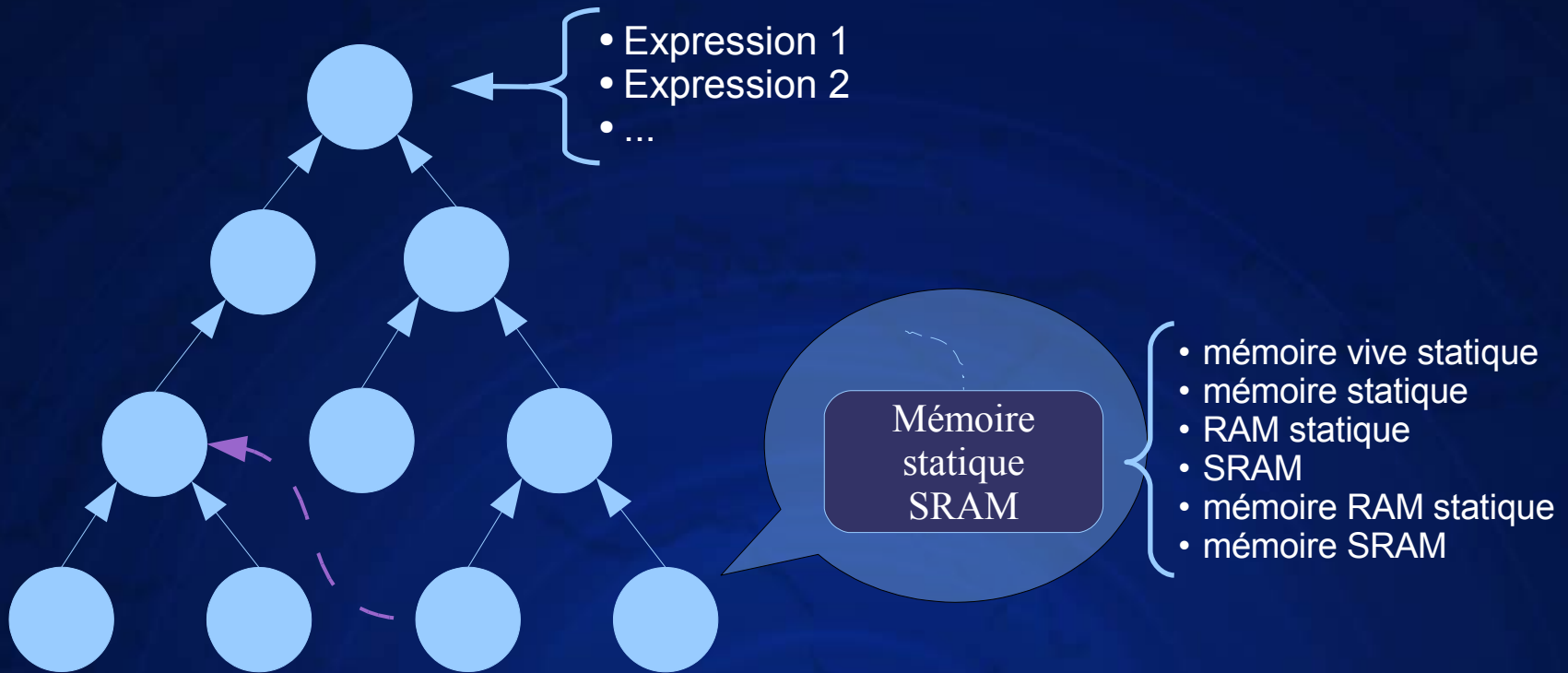
- Site Web des cours du département

- Annuaire, plan cours, pages Web

Représentation graphique de l'ontologie

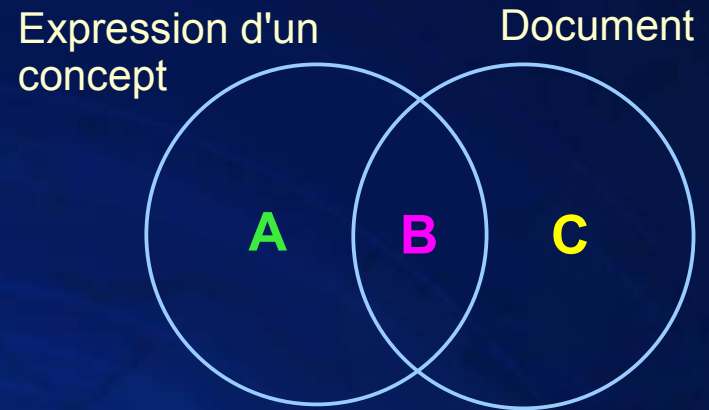


Représentation graphique de l'ontologie



Comment identifier les concepts?

- **Traitement du texte**
 - Correspondance entre les mots*
 - Extraction d'information
- **Approche ensembliste**
 - Pondération de chaque ensemble (A, B, C)

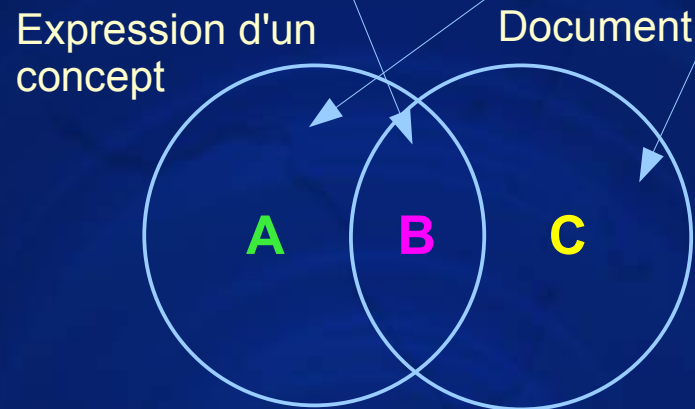


$$\varphi = \alpha A + \beta B + \gamma C$$

Exemple de correspondance

Document : Les langages de programmation actuels...

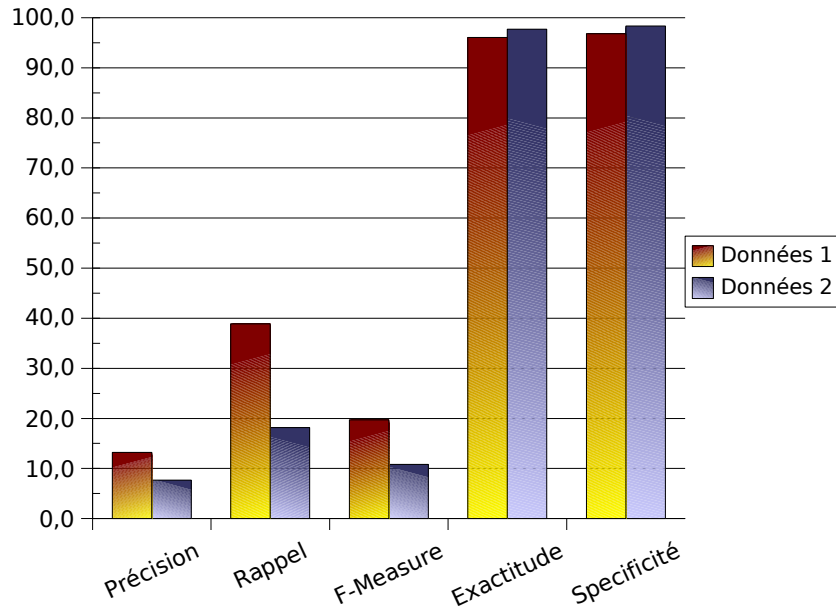
Ontologie : langages de programmation procéduraux



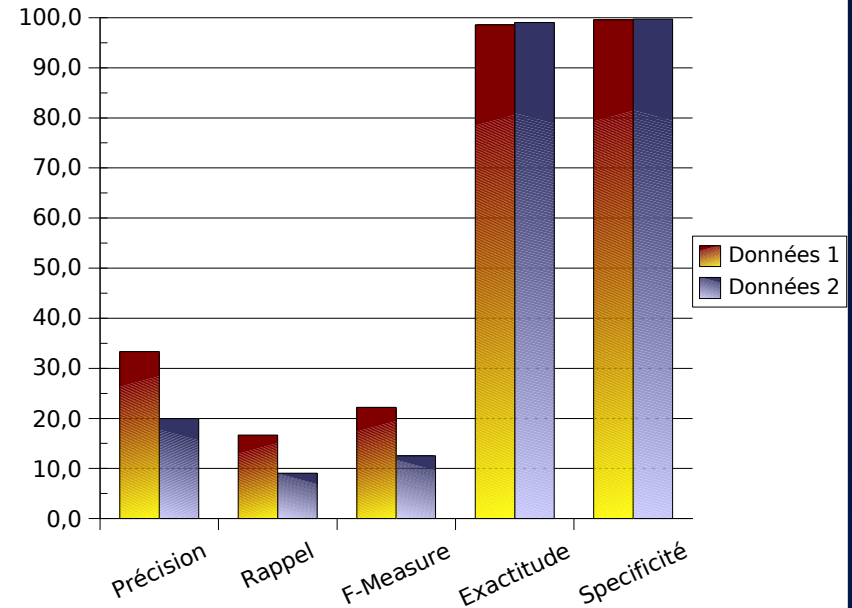
Baseline

Résultats : *baseline* (IE)

Résultats du test de base pour le jeu de pondération A



Résultats du test de base pour le jeu de pondération B



Problèmes avec les mots

- **Synonymie**
 - Plusieurs expressions différentes pour exprimer une même idée
- **Polysémie**
 - Un même mot peut avoir plusieurs *sens*
- **Accords, conjugaisons, etc.**

Problème de correspondance avec des mots

Document : Les langages de programmation actuels...

Ontologie : langage de programmation procédural

Solution aux problèmes linguistiques

- Traitement de la langue naturelle (NLP)
 - Déterminer catégorie lexicale
 - Retourner le lemme
 - Segmenter le texte
 - Identifier les mots

« L'indexation de documents
est complexe. »

Mot	Lemme	Cat. Lex.
L'	<i>le</i>	determinant
indexation	indexation	nom
de	de	de
documents	<i>document</i>	nom
est	<i>être</i>	verbe
complexe	complexe	adjectif

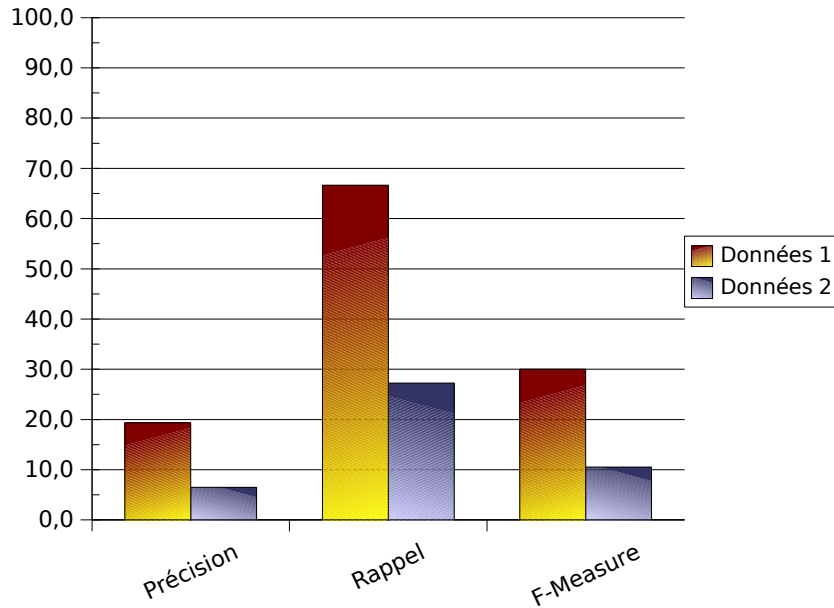
Segmentation

[...] Mon problème de mémoire à court terme me désole. Je ne peux [...] Nous avons reçu les notes des travaux pratiques. Le cours de mécanique statique est difficile. Je n'imagine [...]

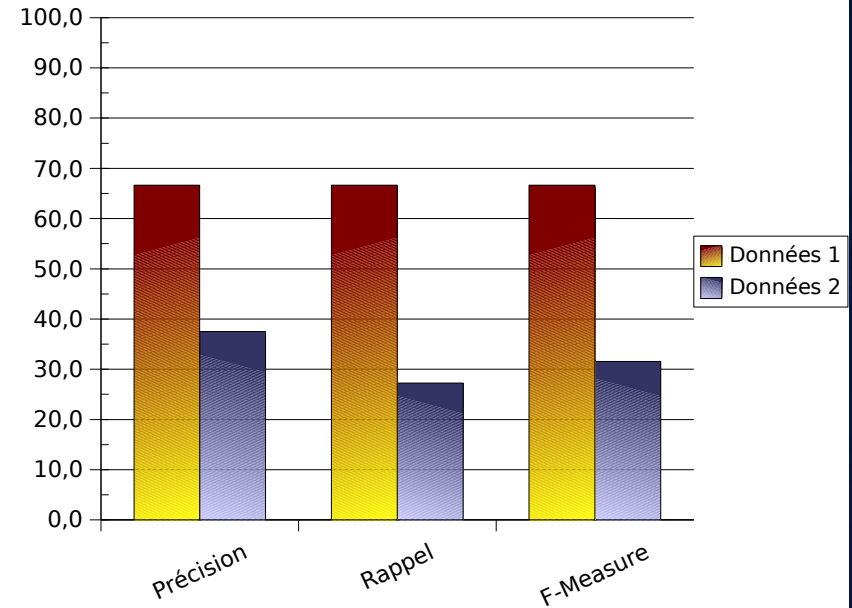
Expression du concepte recherché : « mémoire statique »

Résultats avec traitement linguistique

Résultats du test de base +
segmentation

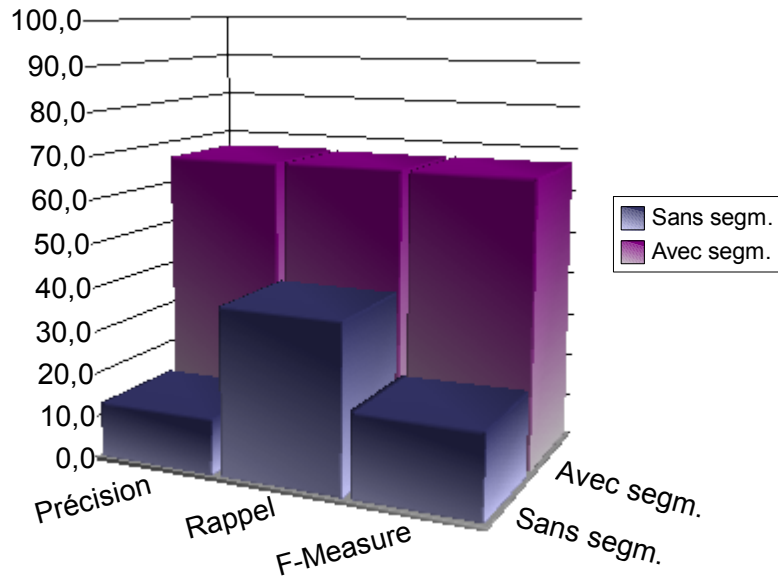


Résultats du test de base +
segmentation + Lex. + filtre

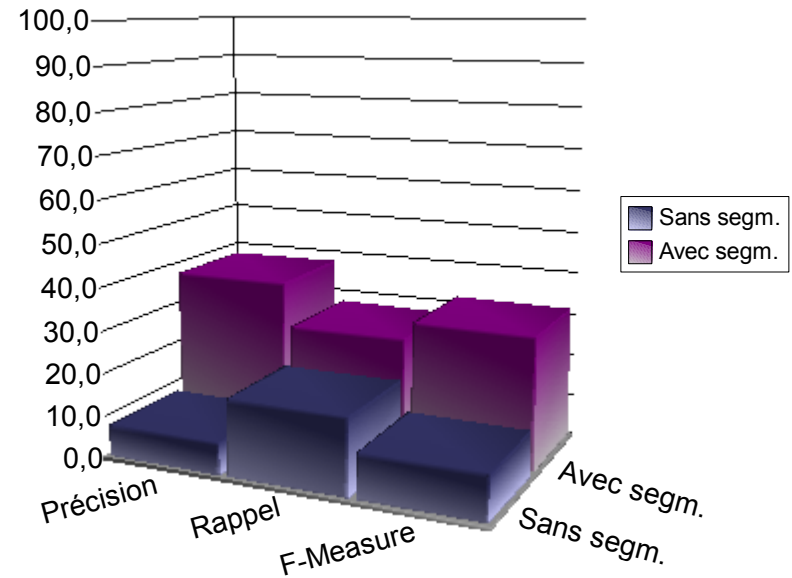


Comparaison avec le traitement linguistique

Comparaison avec et sans traitement linguistique sur données 1



Comparaison avec et sans traitement linguistique sur données 2

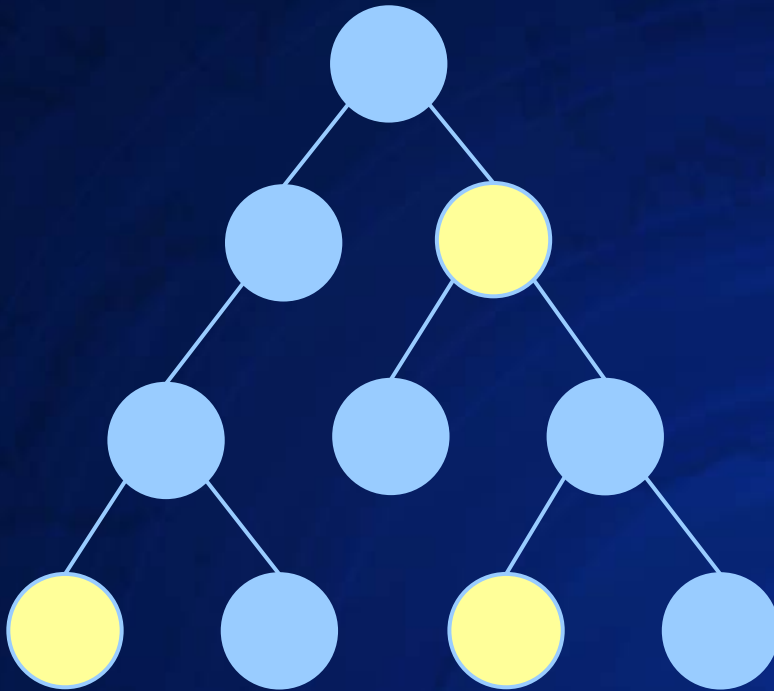


Étapes du processus

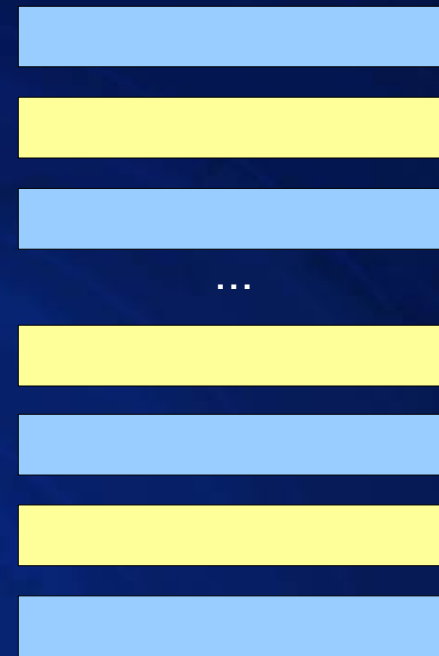
- Extraction d'information
 - Performance : ~20%
- Analyse syntaxique
 - Performance : ~66%
 - Gain : +300%
- ..?

Possibilité de post-traitement?

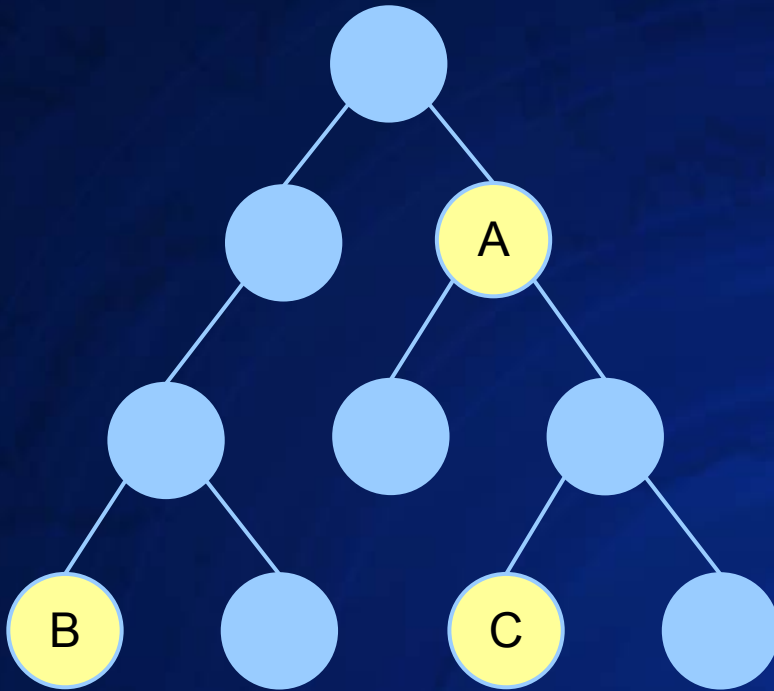
Organisation hiérarchique



Organisation linéaire



Phase de correction

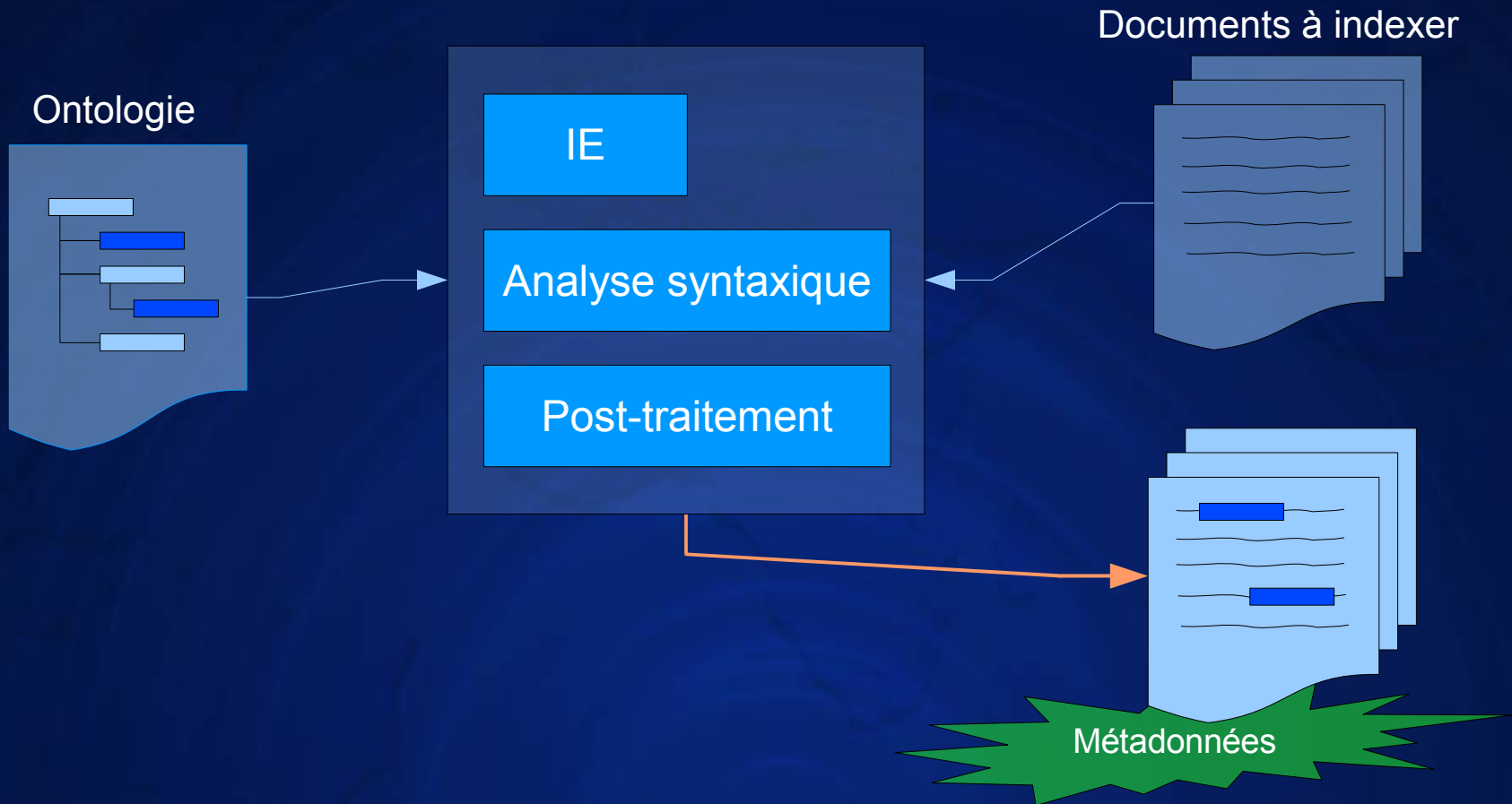


- Possibilités de raffinement
 - Éliminer les réponses subsumées {A}
 - Éliminer les réponses isolées {B}
- Sans connaissance particulière

Plan de la présentation

- Introduction
- Objectifs de recherche
- Présentation du projet
- Conclusion
 - Retour
 - Travaux futurs

Résumé du processus d'indexation



Travaux futurs

- Analyser les résultats
 - Augmenter la taille de notre échantillon
- Améliorer les performances
 - Utiliser la structure des documents (HTML)
 - Exploiter les possibilités de l'analyseur syntaxique
 - Valider la méthode de pondération
 - Apprentissage machine
 - Tirer profit de la structure des données

Conclusions

- Le Web sémantique est un projet ambitieux, mais réalisable et souhaitable!
- Pour y arriver il faut des outils d'indexation sémantique simples
- Simple implique entièrement automatique

Conclusions (suite)

- Le traitement de la langue naturelle (NLP) est un outil précieux
- La manipulation des connaissances demande
 - de nouvelles technologies (ontologie)
 - de nouvelles techniques d'évaluations

Questions..?

