Partage à grande échelle de ressources hétérogènes et réparties :

Localisation de métadonnées

Nicolas.Lumineau@lip6.fr

Laboratoire d'Informatique de Paris VI (LIP6 – Pôle IA)

Séminaire CDS – 12/03/2004

Projet PADOUE

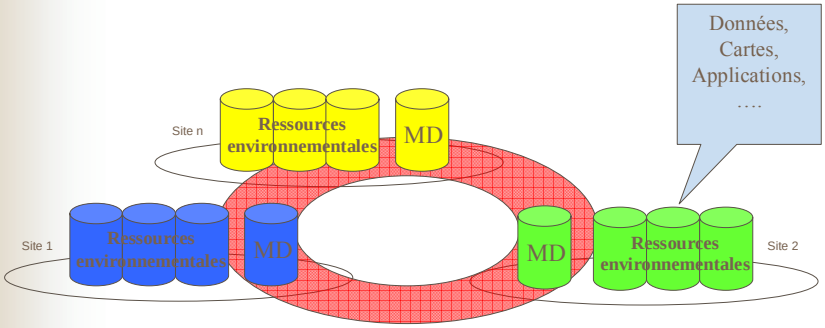
PArtage de DOnnées Utilisées en Environnement

Page web : http://www-poleia.lip6.fr/padoue

Projet pluridisciplinaire de l'ACI GRID : 2002-2005

Contexte

Accumulation de grandes quantités de ressources autonomes et inexploitées



MD: MétaDonnées

Objectifs

- Construire un réseau de partage de métadonnées qui …
  - Passe à l'échelle
  - Gère plusieurs schémas de MDs différents
  - Permet la localisation efficace de métadonnées pertinentes

Plan

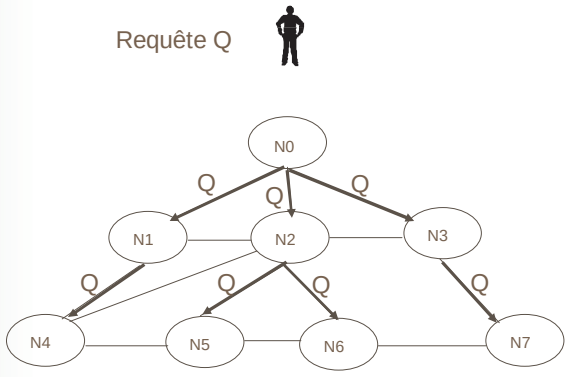
- Choix architecturaux
  - Stockage des Métadonnées
    - Les Systèmes Pair à Pair
  - Accès aux Métadonnées
    - Couplage avec le médiateur «LeSelect»
  - Proposition d'architecture
- Optimisation du processus de localisation
  - La sémantique de PADOUE
  - Connaissance Réseau
    - Système VENISE
  - Connaissance Communautaire
    - Liens intercommunautaires
  - Connaissance des Requêtes
    - Délégation de charge
- Bilan

Les Réseaux Pair à Pair

- Relation d'*égal à égal* entre les nœuds du réseau:
  chaque nœud est à la fois client **et** serveur.

- Caractéristiques principales:
  - Pas de connaissance globale du réseau
  - Pas de coordination globale des noeuds
  - Chaque nœud ne connaît que les nœuds constituant son voisinage
  - Toutes les données sont accessibles à partir de n'importe quel nœud
  - Volatilité des noeuds

Processus de propagation des requêtes

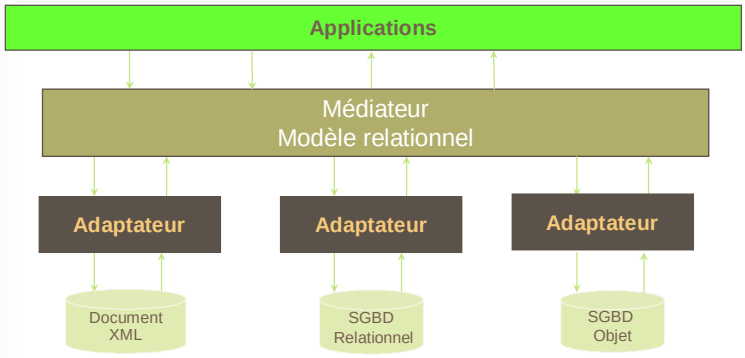


Classification des différents systèmes P2P

- P2P Pur (GNUTELLA)
  - Respect des caractéristiques précédentes
- P2P Hybride (Napster)
  - Données distribuées mais index centralisé
- P2P Structuré (Chord, P-Grid, CAN, …)
  - Index distribué et stocké par DHT (Distributed Hash Tables)
- P2P Hiérarchique (Super-Peer, Kazaa,…)
  - Couplage C/S et P2P
- P2P Sémantique (SON, Routing Indices, …)
  - P2P Pur avec routage enrichi de critères sémantiques

Architecture de médiation

Objectif: Accès transparent aux données hétérogènes



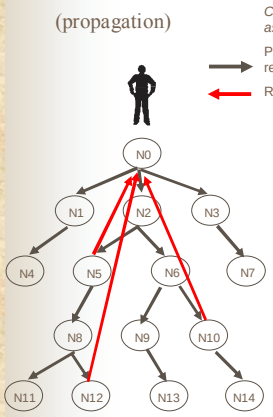
Couplage P2P/médiateur

- Utilisation du médiateur «LeSelect»
  - Diversité des schémas de Métadonnées
- Problématique:
  - La clause FROM des requêtes de médiation
- Proposer une utilisation du médiateur de manière à permettre le processus de propagation des requêtes

Gestion des requêtes

P2P

(propagation)

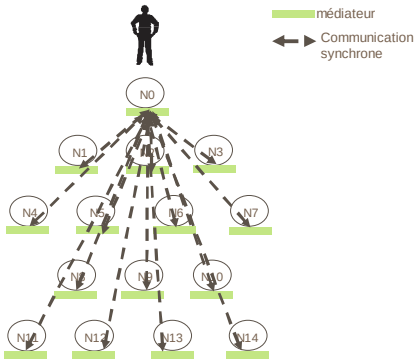


Connaissance des nœuds voisins

Décentralisation de l'interrogation

LeSelect

(interrogation précise)



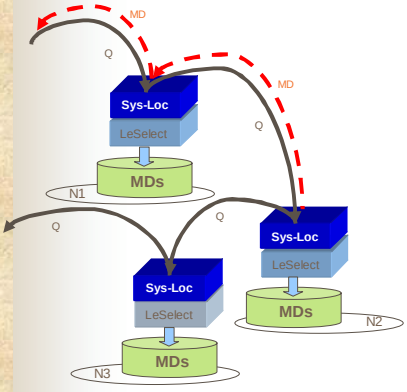
Connaissance de tous les nœuds

Centralisation de l'interrogation

Double utilisation du médiateur

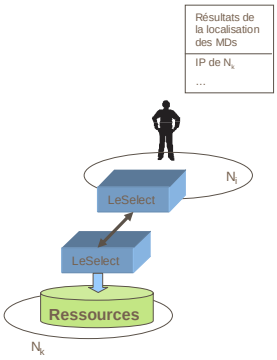
Locale

(accès aux métadonnées)

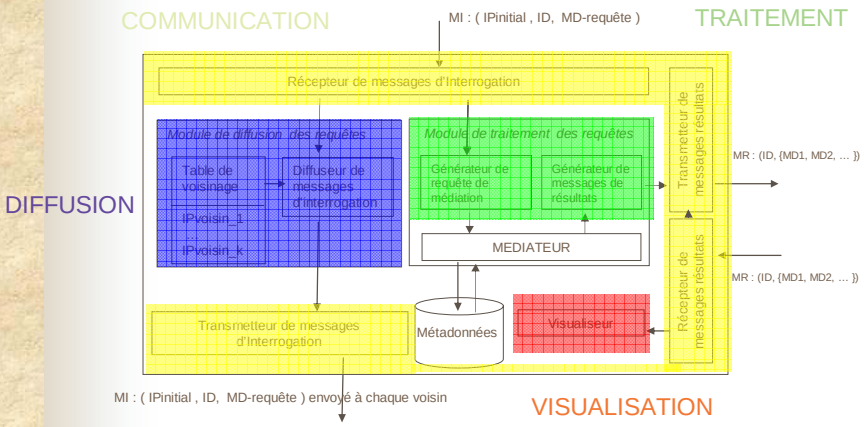


Classique

(accès aux ressources)



Proposition d'architecture



Plan

- Choix architecturaux
  - Stockage des Métadonnées
    - Les Systèmes Pair à Pair
  - Accès aux Métadonnées
    - Couplage avec le médiateur «LeSelect»
  - Proposition d'architecture
- Optimisation du processus de localisation
  - La sémantique de PADOUE
  - Connaissance Réseau
    - Système VENISE
  - Connaissance Communautaire
    - Liens intercommunautaires
  - Connaissance des Requêtes
    - Délégation de charge
- Bilan

La Sémantique de PADOUE (1)

- Connaissance sur le réseau
  - Vecteur Thématique

Thème	"climatologie"	"hydrologie"	"océanologie"	"océanographie"	…
Proportion	**0,35**	**0,25**	**0,40**	**0**	**…**

- Connaissance sur les utilisateurs
  - Vecteur Communauté

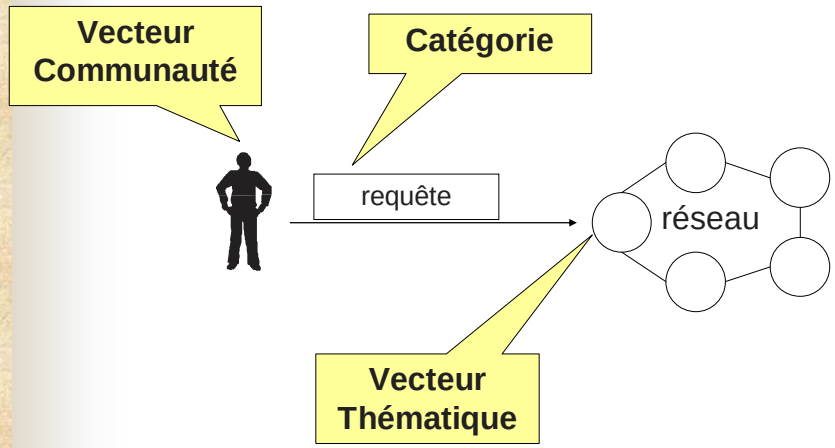
Thème	"climatologie"	"hydrologie"	"océanologie"	"océanographie"	…
Présence	**1**	**0**	**1**	**0**	**…**

La Sémantique de PADOUE (2)

- Connaissance sur les requêtes
  - Catégories de requêtes
    - **Requête Générale (RG)**
      - interrogation du noyau des MDs
    - **Requête Spécifique (RS)**
      - interrogation du noyau + attributs métiers des MDs

	Titre	Date	Climat	Unité	…
	Relevé Pluviométrique	**2003**	**Océanique**	**Cm**	**…**

Bilan des Sémantiques



Complémentarité des approches

- Connaissance Réseau (Vecteur Thématique)
  - ↳ **Organisation du réseau**
- Connaissance Communautaire (Vecteur Communautaire)
  - ↳ **Optimisation du routage des requêtes**
- Connaissance des requêtes (Catégories de requêtes)
  - ↳ **Répartition de la charge des noeuds**

Organisation du réseau

- Objectif:

  Rapprocher logiquement les nœuds sémantiquement proches

- Contrainte:

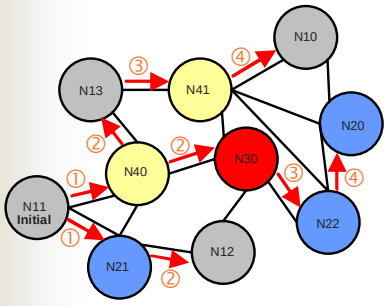
  Pas de connaissance globale du réseau

- Solution:

  Clusterisation du réseau par technique d'apprentissage

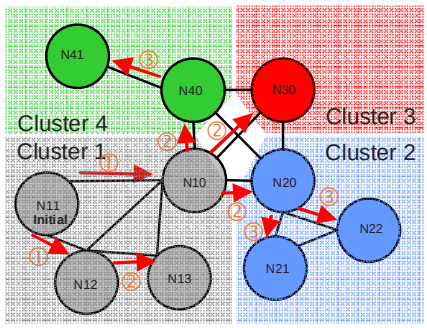
Intérêt de la Clusterisation

- Réduire le nombre de sauts nécessaires à la localisation de métadonnées pertinentes



Réseau NON clusterisé

(Construction aléatoire des voisinages)



Réseau clusterisé

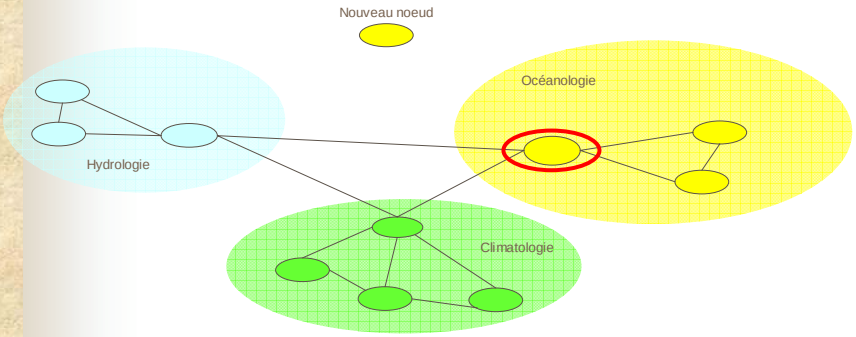
(Construction par sélection des voisins les plus pertinents)

VENISE

serVicE of Node Insertion in Semantic clustEr

- Rôle:

  Service Web qui détermine le nœud du réseau le plus pertinent pour gérer l'insertion d'un nouveau nœud.

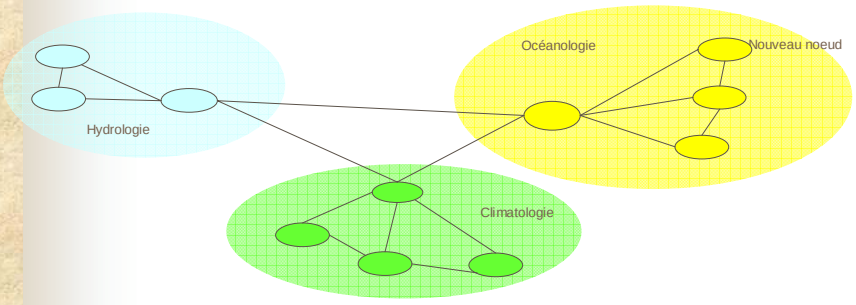


VENISE

serVicE of Node Insertion in Semantic clustEr

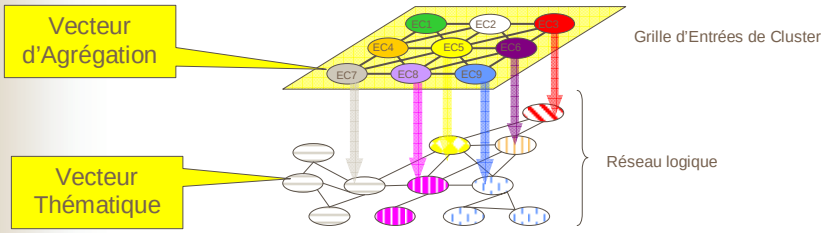
- Rôle:

  Service Web qui détermine le nœud du réseau le plus pertinent pour gérer l'insertion d'un nouveau nœud.



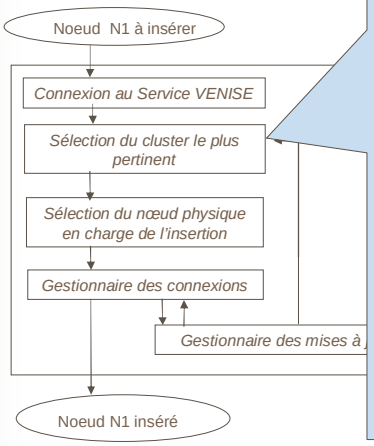
Entrée de Cluster

- Représentation symbolique d'un cluster,
  - stockée par le service web
  - qui pointe sur le nœud du cluster en charge de l'insertion des nouveaux nœuds
  - qui contient une représentation sémantique agrégée de l'ensemble des nœuds du cluster: *Vecteur d'Agrégation*



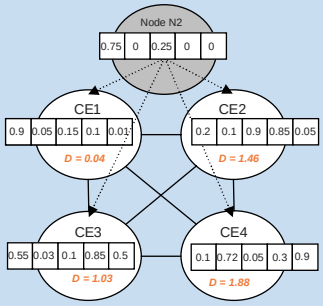
Protocole d'insertion de nœud

- Architecture générale



Calcul de la distance entre le Vecteur Thématique de N1 et le Vecteur d'Agrégation de chaque Entrée de cluster.

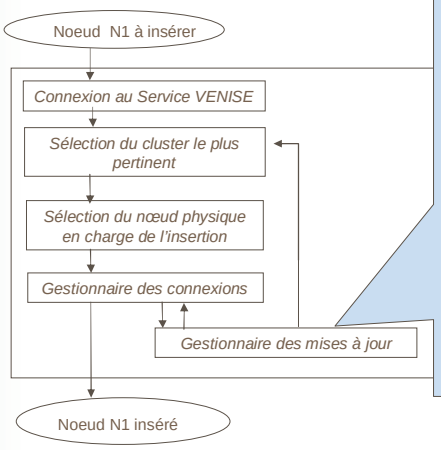
L'entrée élue est celle qui minimise cette distance



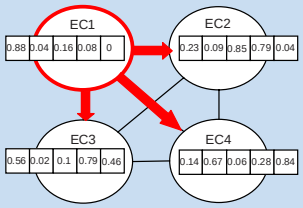
$$EC1 = \arg\min_{e \in E} \| vt_{N1} - va_e \|^2$$

Protocole d'insertion de nœud

- Architecture générale



Mise à jour des Vecteurs d'Agrégation de l'Entrée de Cluster sélectionnée (maintien de la représentation sémantique du cluster) et des Entrées de Cluster voisines (maintien d'une cohérence sémantique inter-clusters).



$$va_e \leftarrow va_e + \alpha.(vt_{N1} - va_e)$$

Maintien des liens physique inter-clusters (pour conserver la connexité du réseau)

Optimisation du routage des requêtes

- Objectif:

  Exploitation des communautés d'utilisateurs

- Contrainte:

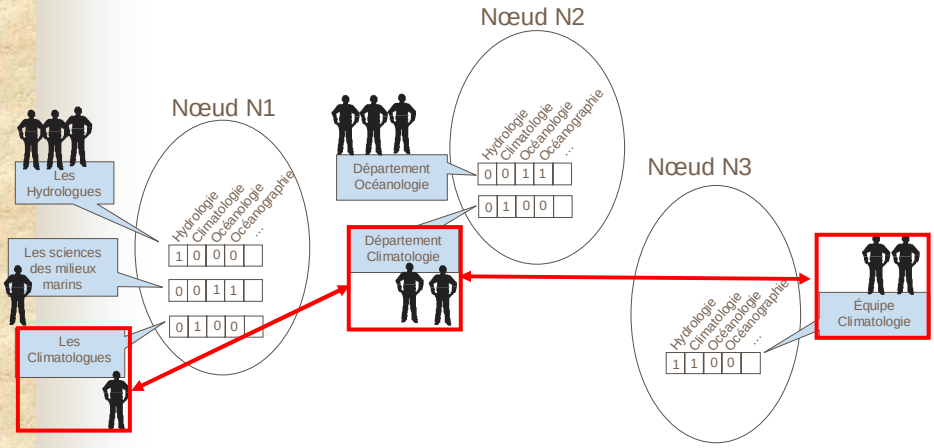
  Pas de définition globale des communautés

- Solution:

  Utilisation des liens entre les communautés distantes de centres d'intérêt communs

Définition des communautés

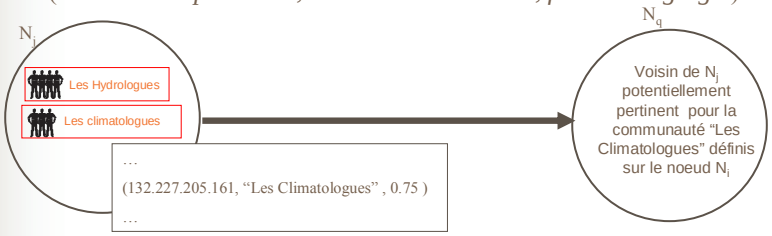
- Définition des *Vecteurs Communautés* sur chaque noeud



Les expériences d'une communauté

- Les Liens de Pertinence basés sur le feedback des membres d'une communauté sur les nœuds filtrant leurs requêtes
- Représentation :

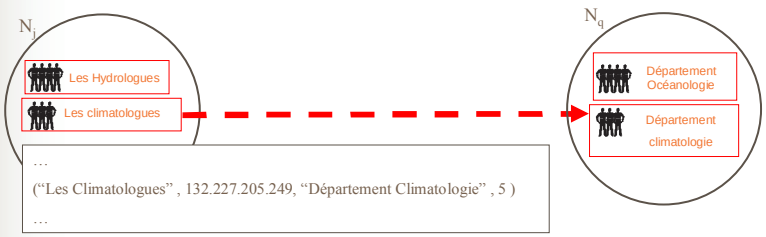
  (*IP du nœud pertinent , Nom de communauté, feedback agrégé*)



La ressemblance entre communautés

- Les Liens intercommunautaires basés sur la similarité entre communautés.
- Représentation :

(Nom communauté locale, *IP du noeud distant , nom de communauté distante, fraîcheur du lien*)



Métriques de comparaison des communauté

- Gestion des liens inter-communautaires:
  - Approche statique

    Par dénombrement des thèmes communs entre deux Vecteurs Communautés

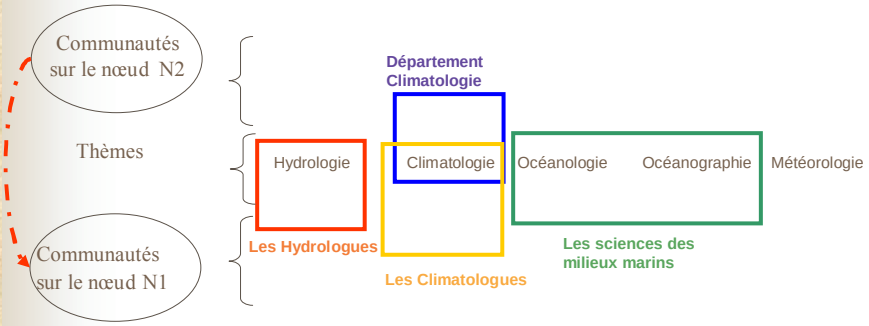
    (def-similarité)
  - Approche dynamique

    Par comparaison des feedbacks agrégés des communautés

    (exp-similarité)

Def-similarité:

Création des liens inter-communautaires



"Département climatologie" et "Les climatologues" sont *def-similaires*

Exp-similarité :

Mise à jour des liens inter-communautaires

Département Climatologie

N1	0,95
N2	0,92
N3	0,89
N4	0,75
N5	0,60
N6	0,25
…	…

Correlation de Pearson : ω

ω("Dept. Clim.", "Les Climatologues")

0,75

ω("Dept. Clim.", "Les Hydrologues")

0,90

ω("Dept. Clim.", "smm")

0,30

Les climatologues

N1	N3	N4	N7	…
0,90	0,92	0,25	0,80	…

Les Hydrologues

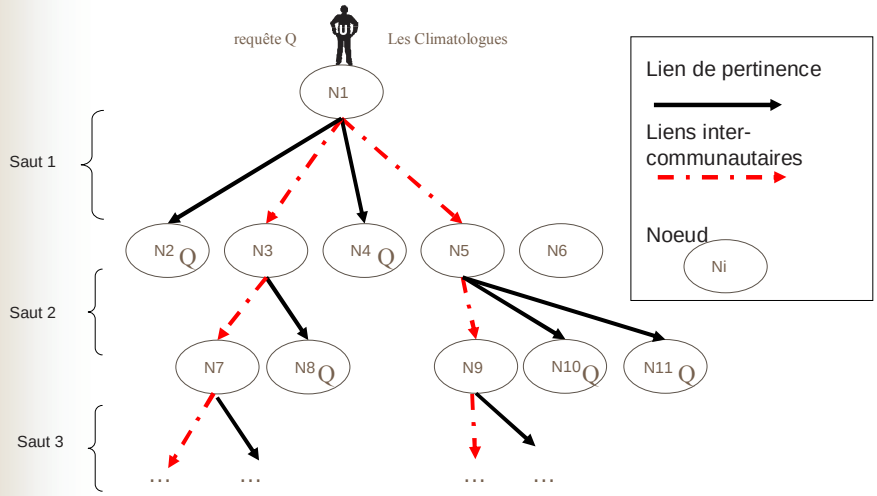
N1	N2	N3	N6	…
0,94	0,96	0,85	0,20	…

Les sciences des milieux marins

N3	N5	N6	N9	…
0,15	0,92	0,95	0,80	…

"Département Climatologie" et "Les Hydrologues" sont *exp-similaires*

GNUCOLLA: Simulateur de propagation des requêtes



Répartition de charge

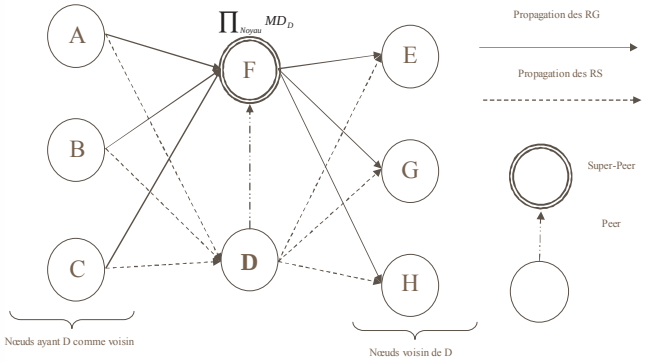
- Objectif:

  Exploiter la structure des MDs et la catégorie des requêtes

- Principe:

  Un nœud peut déléguer le traitement de ses Requêtes Générales à un *Super-Peer* qui contient un répliquât du noyau des métadonnées.

Délégation du traitement des requêtes générales



Bilan & Travaux Futurs

Architecture Pair à Pair intégrant un médiateur pour le partage de métadonnées.

3 axes de recherche pour l'optimisation du processus de propagation des requêtes:

- Clusterisation du réseau basée sur le contenu sémantique des nœuds.
  - Introduction de critères géographique dans le choix des voisins (clusters à multi-entrées)
- Partage de l'expérience des communautés d'utilisateurs pour réduire le nombre de nœuds sur lesquels la requête est exécutée.
  - Utilisation de l'usage pour compenser le manque d'expériences des communautés
- Répartition de charge des requêtes en fonction de leur catégorie.
  - Protocole de négociation pour la délégation de requêtes

http://www-poleia.lip6.fr/padoue

Partenaires

- LIP6 (Université Paris 6)
  - A. Doucet, N. Lumineau, S. Gançarski, B. Defude(INT)
- INRIA (Projet Caravel)
  - E. Simon, JP Matsumoto
- LIRMM (Université de Montpellier)
  - T. Libourel
- Cemagref (Lisc Clermont-Ferrand, UMR 3S Montpellier)
  - G. Bonnet, P. Maurel, A. Miralles
- IRD (Montpellier)
  - J.C.Desconnets, N. Moyroud
- CDS (Strasbourg)
  - F. Genova, A. Schaaff