Recherche d'information sur Internet par algorithmes évolutionnaires

Fabien Picarougne

Laboratoire d'Informatique de l'Université de Tours
Polytech'Tours (Département Informatique)
64, avenue Jean Portalis, 37200 Tours

Plan de l'exposé

1. Problématique de la recherche d'information sur Internet
2. Études et approches existantes
3. Modélisation en problème d'optimisation
4. Résolution par des méta-heuristiques (AG, Fourmis, Tabou)
5. Classification et visualisation des résultats de la recherche
6. Conclusions et Perspectives

Motivations : veille stratégique

- Processus de veille stratégique sur Internet (VSST'2004)
  1. Audit des besoins
  2. Recherche de documents
  3. Traitement des données (ECD, mise à jour, …)
  4. Synthèse et diffusion des résultats
- Contexte de cette thèse (point 2) : fournir des documents à des outils de veille (*Web Mining, Etzioni 1996*)

Difficultés liées au Web Mining

- Du point de vue des données
  - Volume important (8 milliards de documents indexés)
  - Données changeantes (40% du web/mois)
  - Données distribuées (pas de topologie d'organisation des données)
  - Données redondantes (30% sont dupliquées)
  - Données hétérogènes (≠ formats)
  - Mesure de la pertinence des données
- Du point de vue des utilisateurs
  - Spécifier la requête
  - Interpréter la réponse d'un outil de recherche
    - Sélection des résultats
    - Présentation des résultats

Solutions

- Actuellement : utilisation de moteurs de recherche classiques (Google, Altavista, …) ou de méta-moteurs
  - Objectif : rapidité de réponse
    - Utilisation d'un index
    - Pertinence indépendante de la requête
  - Problèmes (mise à jour de l'information, requête pauvre, présentation des résultats)

- Notre approche : diffère d'un moteur de recherche classique
  - Notion de temps différente
  - Évaluation de documents plus approfondie

    → Requête utilisateur riche

Objectifs de la thèse

- Définir un outil de recherche :
  - Requête
    - Définition d'une requête d'interrogation riche

    → Mieux cerner les besoins de l'utilisateur
  - Stratégie de recherche

    → Recherche de l'information pertinente

    → Améliorer le temps de réponse du système
  - Présentation
    - Utilisation d'une classification visuelle des résultats

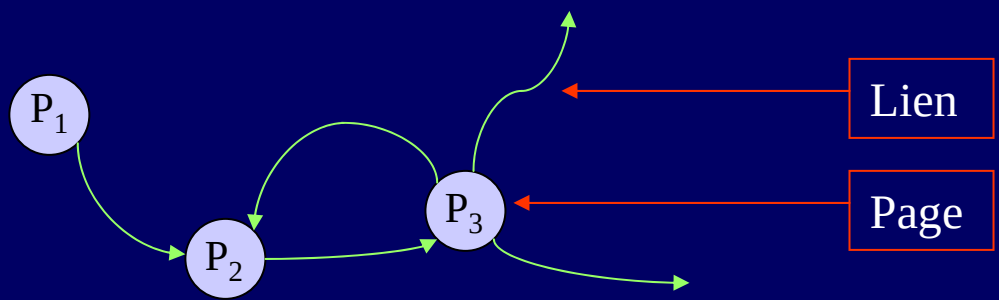
    → Aider l'utilisateur à interpréter les résultats

Contenu

1. Problématique de la recherche d'information sur Internet
2. Études et approches existantes
3. Modélisation en problème d'optimisation
4. Résolution par des méta-heuristiques (AG, Fourmis, Tabou)
5. Classification et visualisation des résultats de la recherche
6. Conclusions et Perspectives

Caractéristiques d'Internet

- Internet vu comme un graphe
  - Caractéristique des documents du web
    - Multimédia
    - Liens hypertextes
  - Modélisation



- Détermination des caractéristiques du graphe

Caractéristiques d'Internet

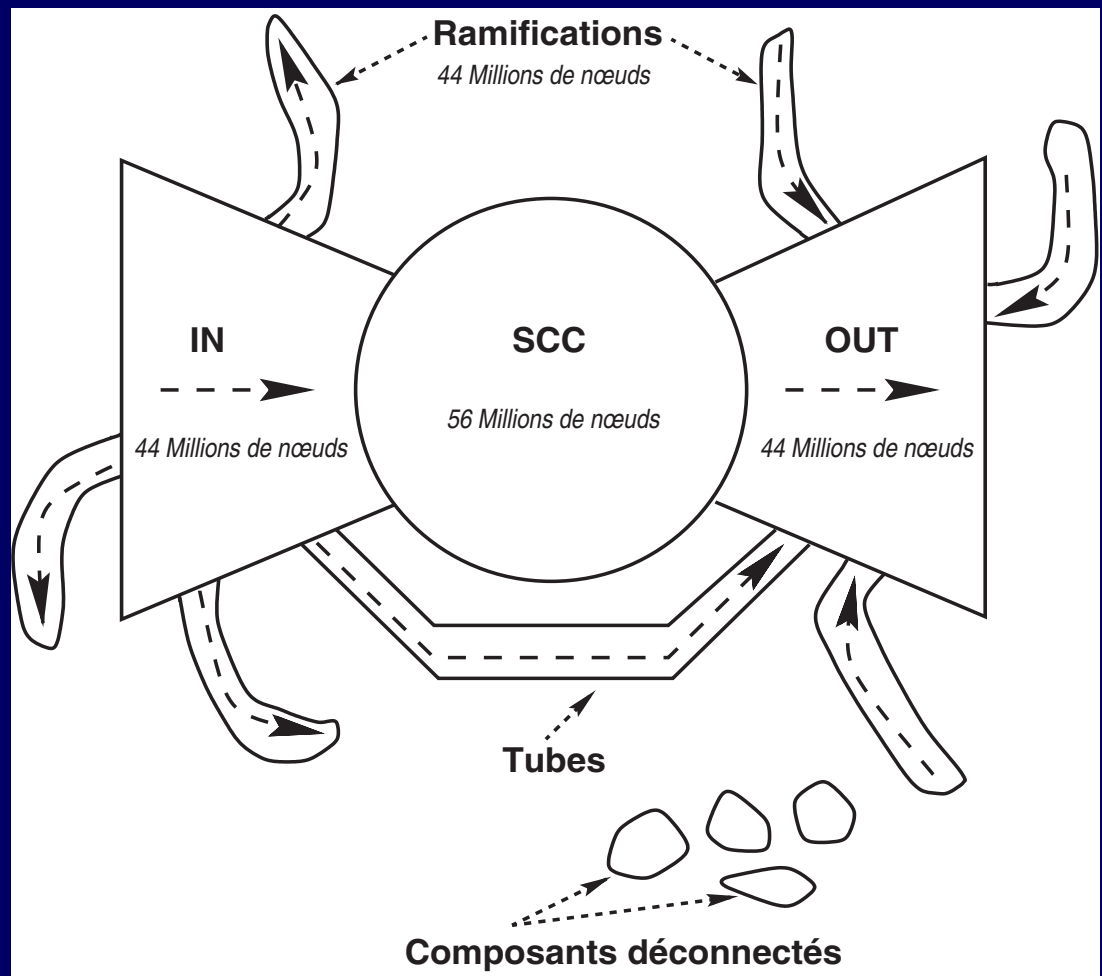
- Topologie du graphe (Nombreuses études)
  - Lois de puissance [Calvert et al 1997] [Faloutsos et al 1999] [Siganos et al 2003]
    - Expressions de la forme $y \propto x^a$
    - Internet : plusieurs niveaux de connexion
    - Caractéristiques des nœuds
      - Nombre de liens sortants proportionnels à l'importance du nœud concerné
      - Les nœuds possédant peu de liens sont plus nombreux que les autres
  - Fractales [Yook et al 2002]
    - Plus un nœud possède d'arcs entrants, plus de nouveaux nœuds auront tendance à se lier à lui (notion d'autorité sur le réseau)

Caractéristiques d'Internet

- Statistiques topologiques
  - [Albert et al 1999] [Barabasi et al 2000]
    - 326 000 documents et 1 470 000 liens
    - Caractérisation du diamètre du web en fonction du nombre de nœuds
    - Extrapolation à 10 milliards de nœuds : 21 liens à suivre en moyenne pour rejoindre n'importe quel point d'Internet
  - [Kleinberg et al 1999] [Kumar et al 2000]
    - Nombre moyen de liens sortants par document : 7,2
  - Dernière étude en date [Broder et al 2000]
    - Utilisation de 200 millions de pages et 1,5 milliard de liens
    - 90% des pages sont fortement interconnectées
    - Résultats : 4 grandes familles de pages

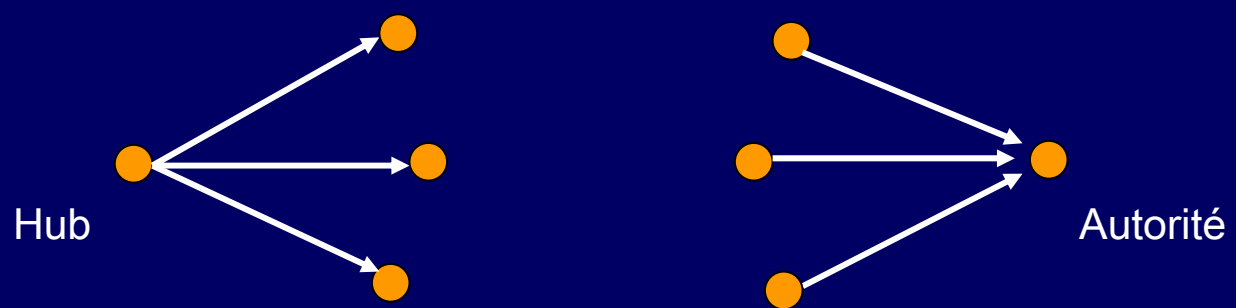
Caractéristiques d'Internet

- Étude de [Broder et al 2000]
- SCC (Strongly Connected Componant)
- Diamètre du graphe :
  - 500 liens maximum
- Diamètre du noyau :
  - 28 liens maximum
  - 16 liens en moyenne



Utilisation dans les moteurs

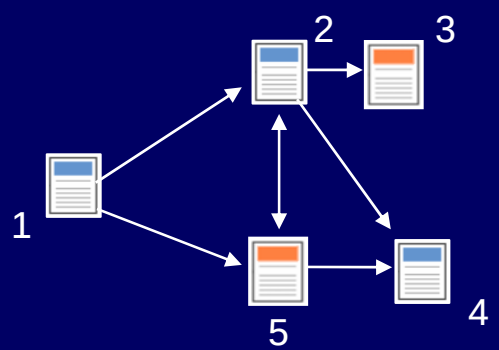
- Mesures de pertinence de pages web
  - Notion d'autorité sur Internet
  - Algorithme HITS [Kleinberg 1999]
    - Utilisation d'un sous-graphe d'Internet
    - Détermination de Hubs et Autorités pour un domaine



- Méthode utilisée dans quelques méta-moteurs

Méthodologie dominante

- Google et le PageRank [Page et Brin 1998]
  - Donne une approximation de la qualité de la page (importance)
  - Ordonne les pages web à partir de l'importance de leurs liens
  - Prends en compte l'ensemble d'Internet
  - Pertinence : probabilité de visite d'un internaute aléatoire



- Ordre de pertinence : 4, 2, 5, 3 et 1

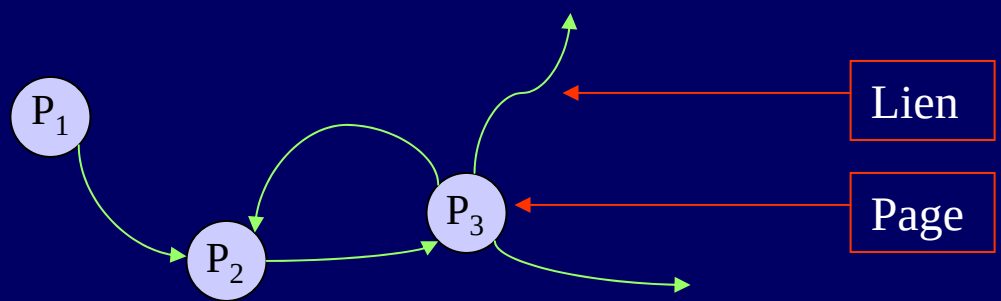
  - Tri de pertinence indépendant de la requête

Contenu

1. Problématique de la recherche d'information sur Internet
2. Études et approches existantes
3. Modélisation en problème d'optimisation
4. Résolution par des méta-heuristiques (AG, Fourmis, Tabou)
5. Classification et visualisation des résultats de la recherche
6. Conclusions et Perspectives

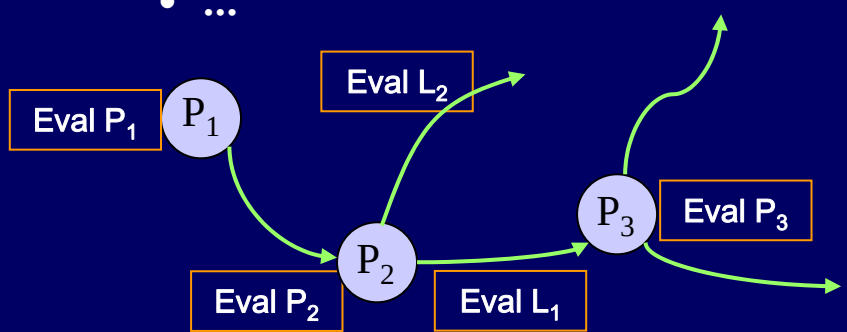
Modélisation

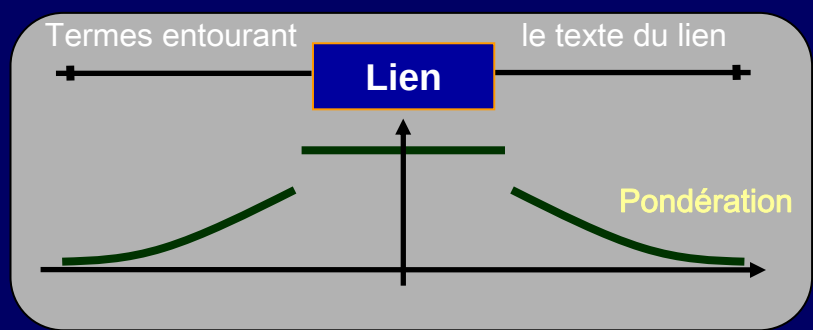
- Optimisation combinatoire :
  - Notion de voisinage pour chaque point de l'espace
  - Établissement d'une fonction d'évaluation
- Codage du problème
  - Espace de recherche : Internet
  - Documents : texte + liens hypertextes
  - Modélisation sous forme de graphe orienté



Modélisation

- Fonction d'évaluation
  - Basée sur la requête utilisateur
  - Plusieurs listes de mots clés (K, S, SN, M, MN)
  - Mesure de la proximité entre les mots clés
  - Utilisation de plusieurs critères
    - Maximiser la présence d'images, de sons, …
    - Favoriser des proportions identiques de mots clés
    - Évaluation des liens hypertextes (présence de mots clés)
    - …





Modélisation

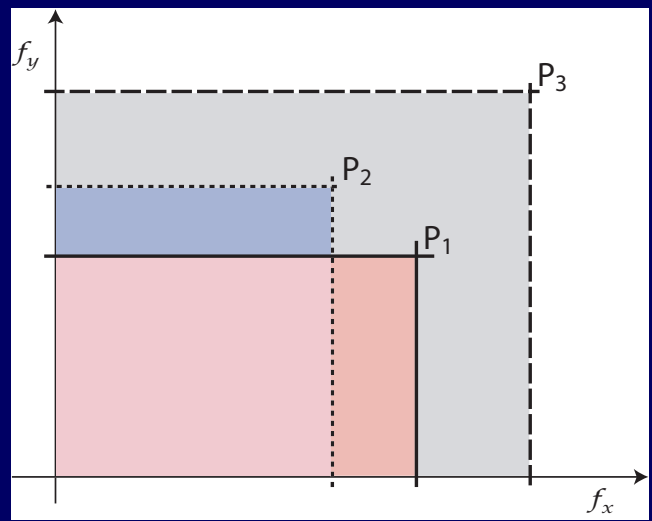
- Évaluation multicritères d'une page S
- $f : S \rightarrow \Re^n$

  n: nombre de critères

- Mesure de la dominance d'une page
  - Valeur booléenne
  - Comparaison : concept de Pareto optimalité

Modélisation

- Exemple
  - Fonction à 2 critères (f_x et f_y)
  - 3 points
  - P_3 domine P_1 et P_2
  - Incertitude entre P_1 et P_2



- En cas d'incertitude, l'utilisateur détermine le poids (l'importance) de chaque critère
- Dominance : somme des poids des critères dominants pour chaque document

Modélisation

→ Modélisation d'un espace de recherche mathématique

- Exploration de l'espace de recherche : Opérateurs de recherche
- 2 types d'opérateurs
  - Opérateur de création heuristique : O_{rand}
    - Utilisation de résultats de moteurs de recherche classiques
  - Opérateur d'exploration locale : O_{explo}
    - Suivi de liens à partir d'une page donnée
    - Utilisation de l'évaluation des liens (texte servant d'ancrage)

Contenu

1. Problématique de la recherche d'information sur Internet
2. Études et approches existantes
3. Modélisation en problème d'optimisation
4. Résolution par des méta-heuristiques (AG, Fourmis, Tabou)
5. Classification et visualisation des résultats de la recherche
6. Conclusions et Perspectives

Méta-heuristiques de résolution

- Pourquoi utiliser des méta-heuristiques ?
  - Graphe d'une très grande taille et dynamique
  - Étude topologique : un agent bien guidé peut trouver n'importe quelle information en suivant efficacement les liens hypertextes [Albert et al 1999]
  - Approches existantes prometteuses : InfoSpiders [Menczer 1999]
- Quelle est la meilleure stratégie de parcours du graphe ?
  - Caractéristiques de l'espace de recherche mathématique difficile à évaluer (stage de DEA)
  - Test de plusieurs types de méthodes
    - Algorithme Génétique
    - Algorithme à base de population d'agents
    - Recherche tabou

Algorithme Génétique

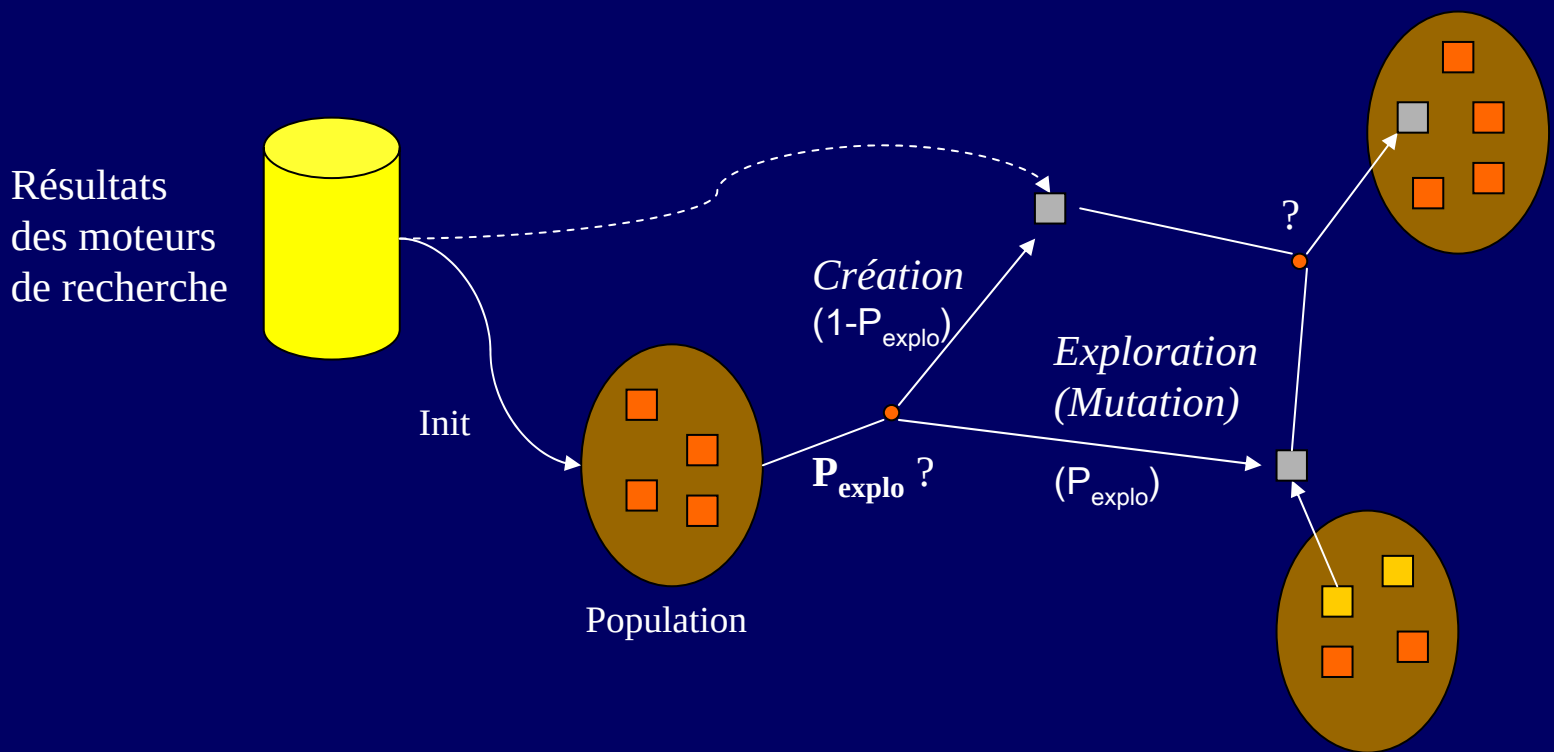
- Principe des algorithmes génétiques [Holland 75]
  - Inspiré des principes biologiques
    - Évolutionnaires (Darwin) : pression sélective du milieu
    - Génétiques : hérédité
  - Modélisation mathématique
    - Population d'individus
    - Codage d'un individu par un chromosome
    - Génération de nouveaux individus
      - Utilisation d'opérateurs génétiques (croisement, mutation)
    - Évaluation (fitness) et insertion des individus dans la population
    - Élimination des individus les plus faibles

Algorithme Génétique

- Application des principes de l'EA avec des opérateurs adaptés aux structures manipulées
- Modélisation génétique du problème
  - Individu = Document
  - Opérateurs génétiques
    - Croisement n'a pas de sens dans cette modélisation
    - Création : O_{rand}
    - Exploration (Mutation) : tournoi binaire + exploration du meilleur lien $\rightarrow$ O_{explo}
  - Évaluation de la page $\rightarrow$ fitness de l'individu

Geniminer

- Algorithme général



Antsearch

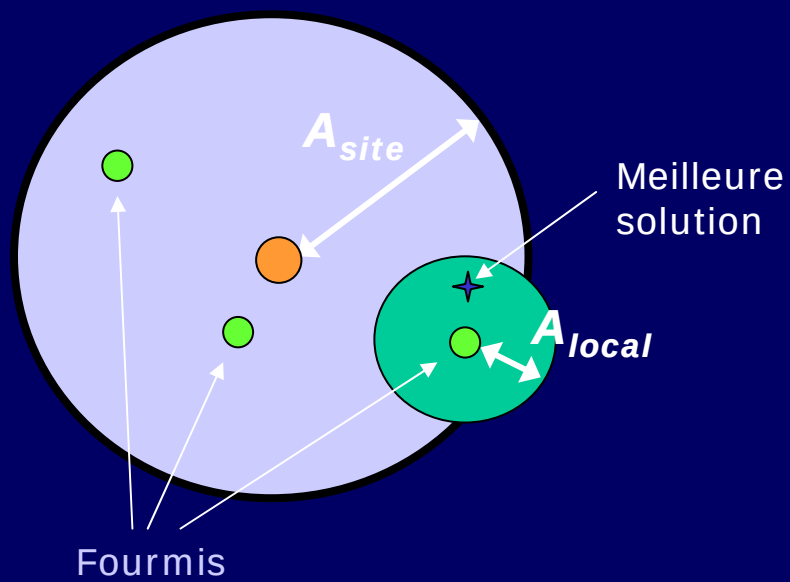
- Adaptation de l'algorithme API [Monmarché 00]
- S'inspire de la stratégie de fourragement des fourmis de la famille *Pachycondyla apicalis*.

- Idée globale
  - Sélection d'un site de chasse près du nid (point de l'espace de recherche)
  - Capture des proies (les documents) dans le site de chasse
  - Périodiquement : changement de place du nid

    → Exploration de l'espace de recherche

Antsearch

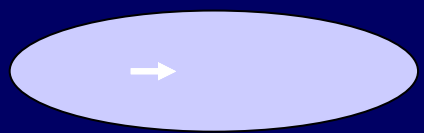
- Algorithme global
  - Nid, site de chasse = 1 page web



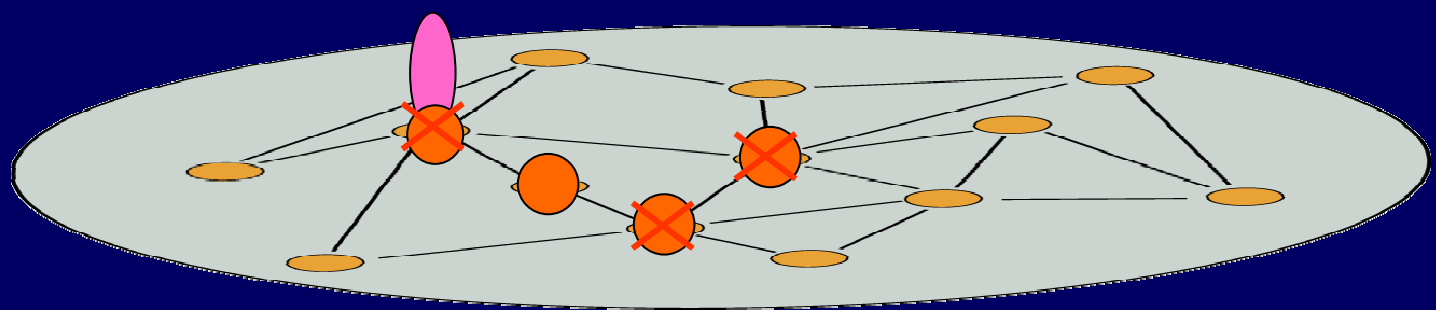
- 2 opérateurs
  - *Nouveau point : O_{rand}*
  - *Exploration locale sur une profondeur donnée : O_{explo}*
- Déplacement du nid
  - Vers la meilleure solution
  - En utilisant l'opérateur O_{rand}

Algorithme tabou [Glover 86]

- Objectif :
  - Utiliser une méthode d'ascension locale améliorée



Liste tabou

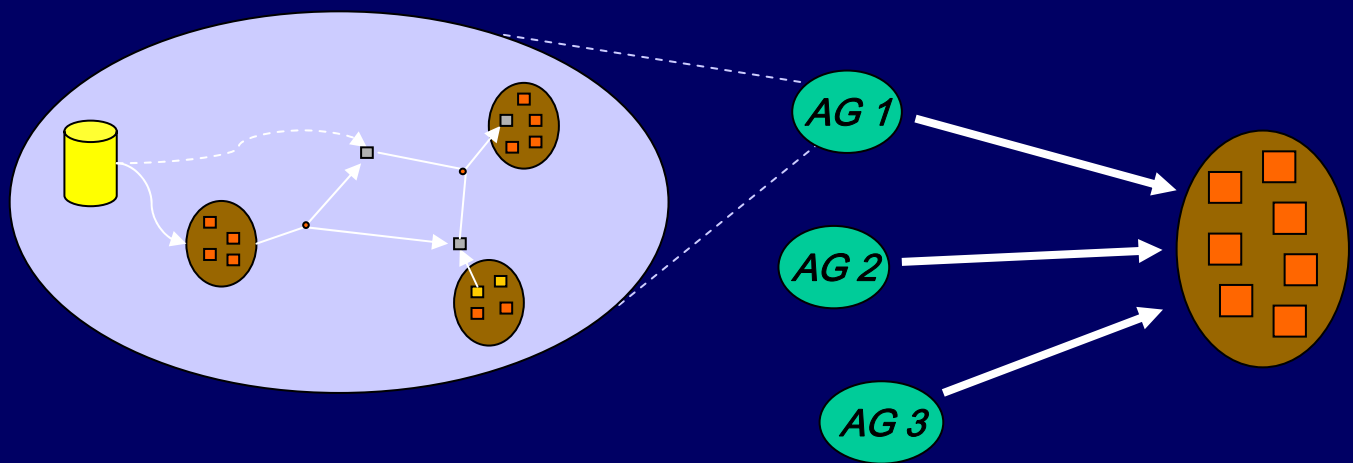


Algorithme tabou : Tabusearch

- Initialisation par l'opérateur O_{rand}
- Guidage de la méthode par l'opérateur O_{explo}
  - Sélection des liens grâce à la fonction d'évaluation
- Arrêt de la recherche sur un nœud terminal (page sans lien)
  - Reprise par initialisation avec l'opérateur O_{rand}

Parallélisation : GeniminerII

- Problèmes des algorithmes séquentiels
  - Lenteur d'exécution :
    - 1 requête (1000 pages) -> 3-4 heures (PC standard)
- Parallélisation de l'AG [Cantú-Paz 2000]
  - Population centralisée
  - Plusieurs AG manipulent la population simultanément



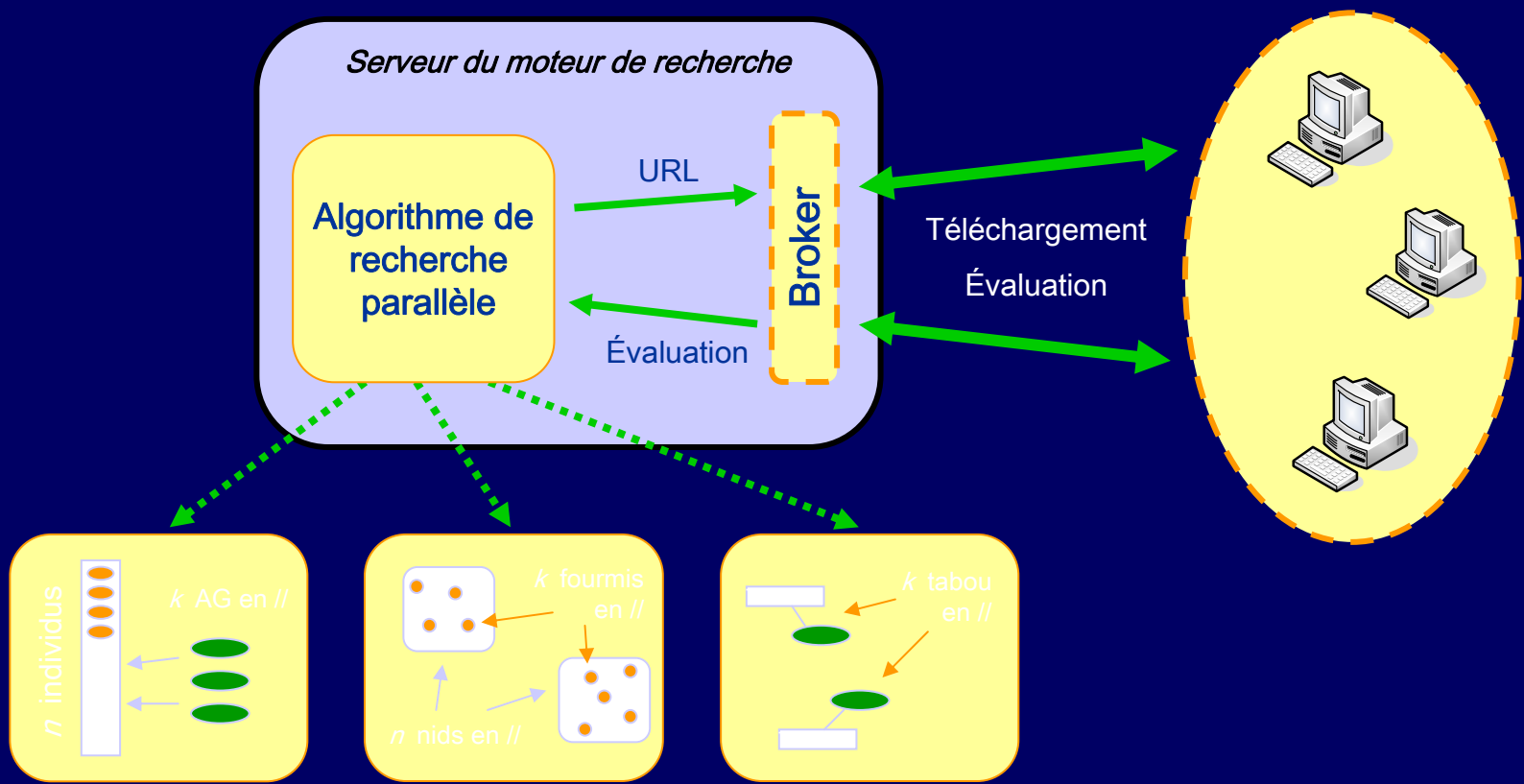
Parallélisation des autres méthodes

- Antsearch
  - Simulation de plusieurs nids en parallèle
  - Les fourmis d'un nid s'exécutent de manière concurrente

- Tabusearch
  - Plusieurs algorithmes tabou en parallèle
  - Interdiction de revenir sur les points fournis par l'opérateur O_{rand}

Généralisation de la parallélisation

- Téléchargement et évaluation des documents sur des clients distants

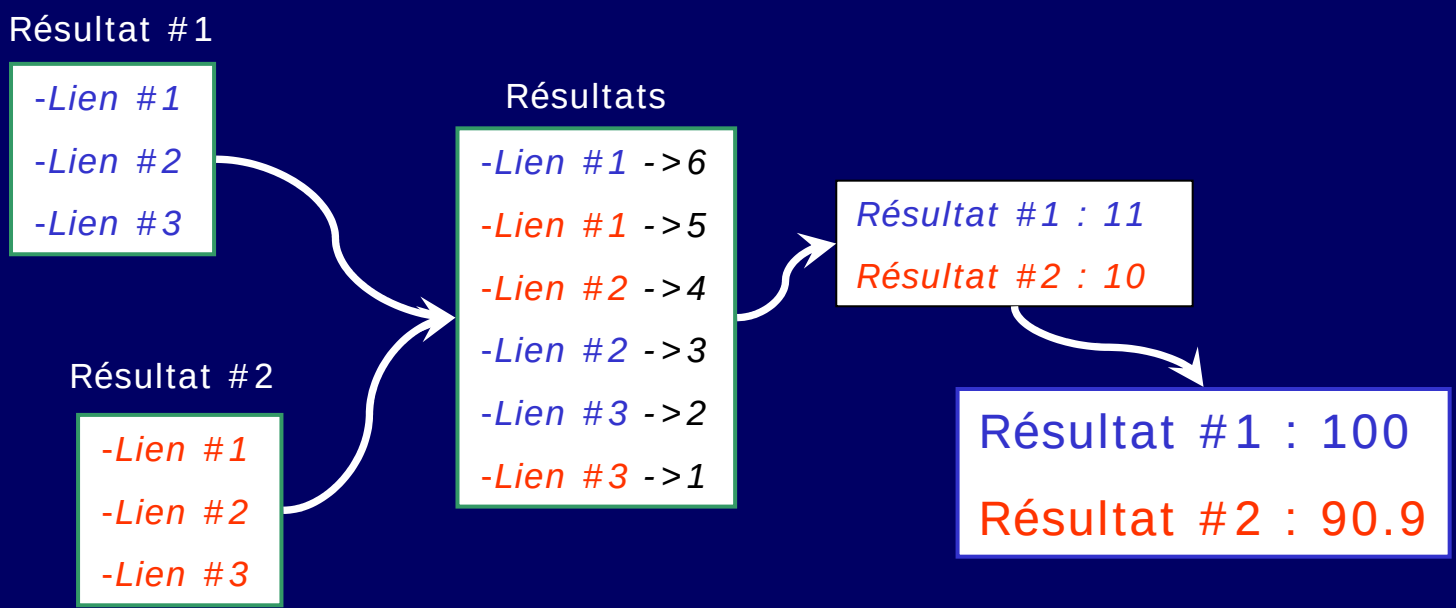


Expérimentations : méthodologie

- Utilisation de la fonction d'évaluation
- Prend en compte la position des réponses dans les résultats
- But : déterminer quelle méthode donne les meilleures pages dans le temps le plus court possible pour l'utilisateur

Expérimentations : méthodologie

- Évaluation des résultats

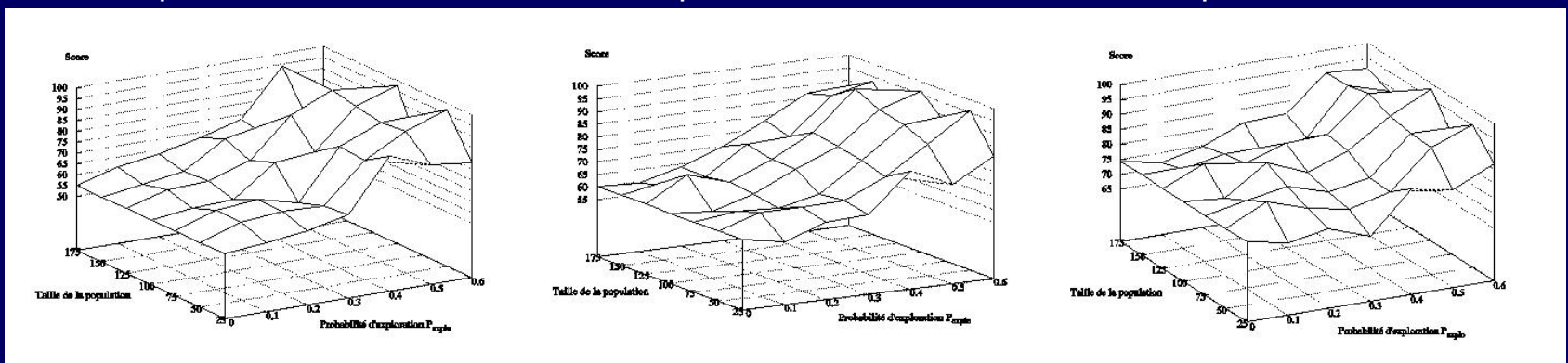


- Score d'un résultat : somme de ses liens

Étude de paramètres

- Protocole :
  - Tests pour 1000 téléchargements
  - Évaluation combinée des différents paramètres
  - Prise en considération de la répartition des résultats dans les listes retournées
- Exemple : GeniminerII

30 premiers résultats | 60 premiers résultats | 100 premiers résultats



Étude de paramètres

- Résultats
  - Pour chaque requête, analyse des 30, 60 et 100 premiers résultats obtenus
  - Pour chaque combinaison de paramètres : moyennes sur 3 exécutions
  - 196 combinaisons de paramètres pour GeniminerII
    - P_{explo} = 0.4
    - Taille de la population = 100
  - 144 combinaisons de paramètres pour Antsearch
    - Faible exploration locale
    - P_{create} = 0.5
    - Peu de fourmis par nid et beaucoup de nids en parallèle
  - 96 combinaisons de paramètres pour Tabusearch
    - 20 listes tabous en parallèle
    - Taille de la liste : 40 places

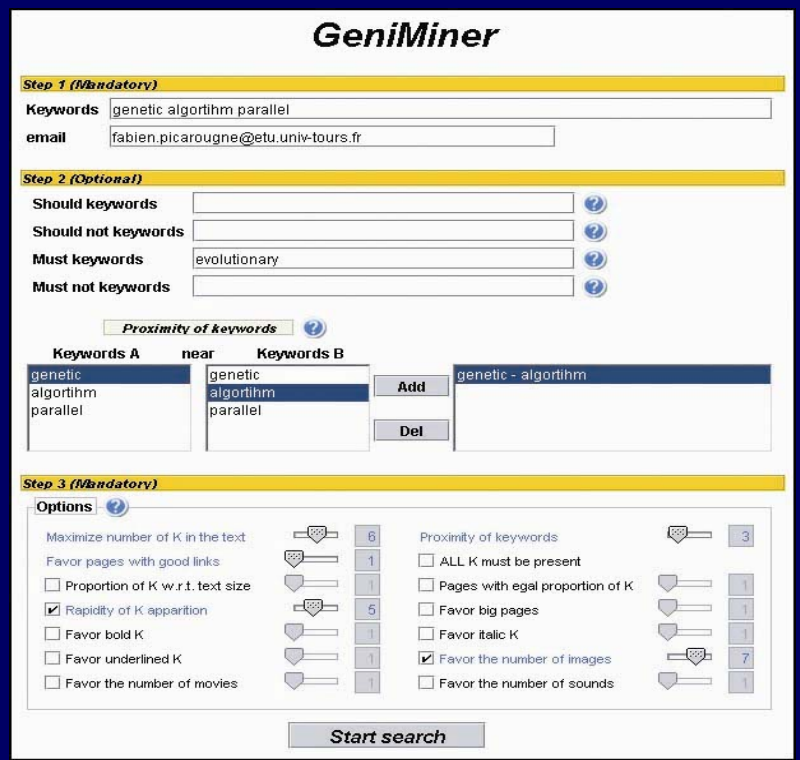
Étude comparative

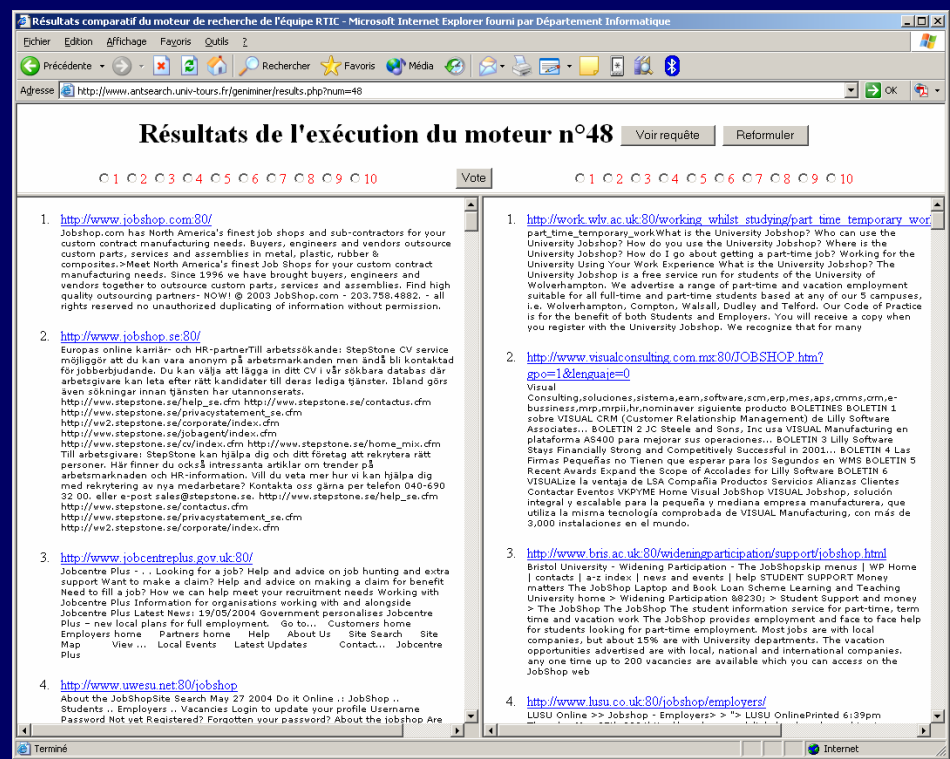
n°	Méta-recherche	GeniminerII		Antsearch		Tabusearch	
		moy	σ	moy	σ	moy	σ
1	81.37	69.99	[9.76]	▶ 89.60	[7.41]	87.43	[6.31]
2	88.36	▶ 88.99	[7.63]	38.87	[4.38]	51.19	[3.79]
3	74.29	▶ 92.80	[4.06]	54.28	[3.63]	55.13	[2.76]
4	▶ 100.00	80.04	[9.41]	79.08	[6.12]	62.86	[12.03]
5	85.64	▶ 87.93	[8.17]	46.06	[3.99]	38.05	[11.40]
6	▶ 100.00	88.45	[7.11]	52.23	[8.38]	62.65	[4.52]
7	▶ 99.27	97.71	[1.69]	66.67	[4.22]	71.95	[2.93]
8	▶ 100.00	72.62	[6.15]	30.72	[5.28]	50.95	[2.95]
9	67.56	▶ 91.71	[4.70]	55.18	[5.47]	50.86	[4.99]
10	▶ 100.00	93.37	[4.87]	58.20	[3.36]	58.39	[4.72]

- L'évaluation caractérise la méthode optimisant le mieux la fonction d'évaluation choisie
- Correspondance fonction d'évaluation – pertinence des documents aux yeux de l'expert ?

Évaluation par des utilisateurs

- Protocole de test





Évaluation par des utilisateurs

- Résultats

Vote	GeniMinerII	Méta-moteur	Égalité
Toutes les requêtes	22 (44,00%)	22 (44,00%)	8 (16,00%)
Requêtes complexes	13 (48,15%)	9 (33,3%)	5 (18,52%)

- Problèmes :
  - Temps d'apprentissage de l'interface
  - Visualisation des résultats

Contenu

1. Problématique de la recherche d'information sur Internet
2. Études et approches existantes
3. Modélisation en problème d'optimisation
4. Résolution par des méta-heuristiques (AG, Fourmis, Tabou)
5. Classification et visualisation des résultats de la recherche
6. Conclusions et Perspectives

Classification des résultats

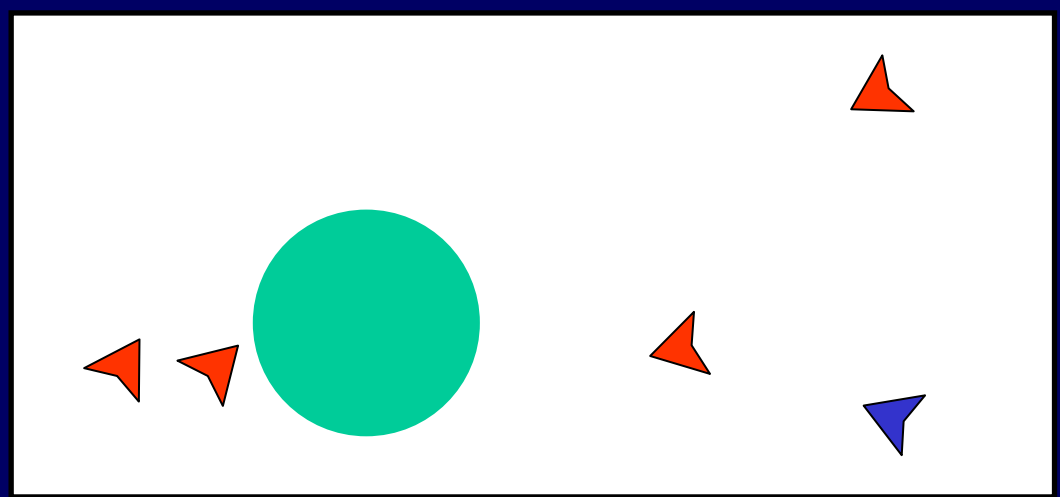
- Classification par nuages d'agents [Proctor et Winter 1998]
  - Inspiration du comportement des oiseaux
  - Intelligence en essaim [Bonabeau et al 1999]
    - Règles locales de coordination des déplacements
    - Comportements et mouvements globaux complexes
- Application à la classification
  - Un agent = un document
  - Déplacement dépend de la similarité des agents dans le voisinage d'un individu

Classification des résultats

- Représentation vectorielle d'un document : un vecteur de poids de mots
  - Poids : méthode TF*IDF [Salton et Buckley 1988]

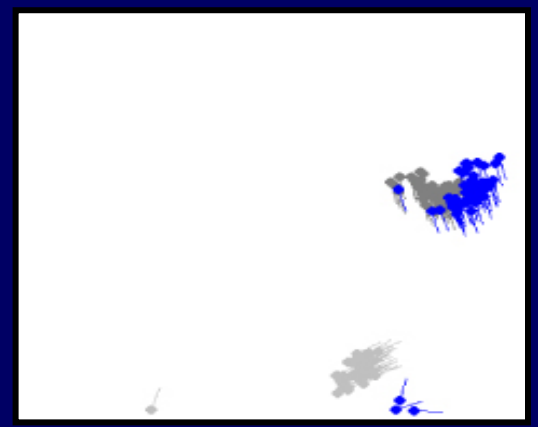
  → Caractérise les mots significatifs d'un document

  - Similarité multilingue : mesure cosinus entre les vecteurs de termes

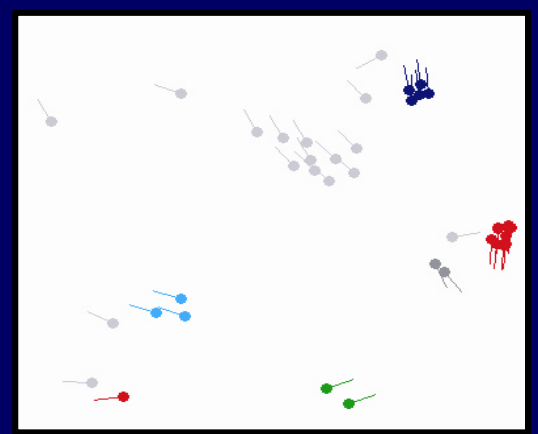


Classification des résultats

- Résultats
  - Tests sur des bases numériques classiques (Machine learning repository : Iris, Soybean, …)
    - Validation de l'algorithme
    - Bonnes performances (nombre de classes trouvées, erreur)



  - Sur GeniminerII : mesure de pertinence humaine



Requête	*Nb classes réelles*	*Nb classes trouvées*	*pureté*
1	3	3	0,84
2	4	4	0,96
3	5	6	0,94

Contenu

1. Problématique de la recherche d'information sur Internet
2. Études et approches existantes
3. Modélisation en problème d'optimisation
4. Résolution par des méta-heuristiques (AG, Fourmis, Tabou)
5. Classification et visualisation des résultats de la recherche
6. Conclusions et Perspectives

Conclusions

- Étude du problème de la recherche d'information sur Internet
  - Contexte de la vieille stratégique
  - Transformation du problème en un problème d'optimisation
- Conception de plusieurs algorithmes de nature différente
  - Comparaison des méthodes
  - → Approche génétique : meilleure méthode
  - Outil complémentaire aux méta-moteurs
    - Meilleur après un apprentissage de l'interface d'interrogation
- Réalisation d'un nouveau système de classification de résultats

Perspectives

- Proposer de nouveaux critères de recherche
  - Interrogation par documents exemples
  - Relevance feedback (requêtes interactives)
- Développer l'aspect veille stratégique
  - Filtrage collaboratif
  - Incrémentalité de la recherche (capitalisation des connaissances)
- Visualisation des résultats (fait l'objet d'une thèse)
  - Ajout d'interactivité dans la classification
  - Utilisation de techniques à base de forces et ressorts

Merci de votre attention…

.....