# DIC9410 - Projet de recherche

## *Agents de Filtrage de l'Information*

**Guy DESJARDINS** UQÀM

**Septembre 2003**

# Plan de la présentation

- **Concept d'Agent logiciel**
- **Problématique**
- **Objectifs du projet**
- **Modèles de filtrage**
- **Méthodologie**
- **Références**

# Plan de la présentation

- **Concept d'Agent logiciel**
  - **Définitions**
  - **Dimensions**
  - **Typologie**
  - **Architectures**
- **Problématique**
- **Objectifs du projet**
- **Modèles de filtrage**
- **Méthodologie**
- **Références**

# Plan de la présentation

- Concept d'Agent logiciel
- **Problématique**
  - **Filtrage**
  - **Modèles**
- Objectifs du projet
- Modèles de filtrage
- Méthodologie
- Références

# Plan de la présentation

- Concept d'Agent logiciel
- Problématique
- **Objectifs du projet**
- Modèles de filtrage
- Méthodologie
- Références

# Plan de la présentation

- Concept d'Agent logiciel
- Problématique
- Objectifs du projet
- **Modèles de filtrage** →
  - Vectoriel classique
  - Booléen étendu
  - Ensembliste
  - Ensembles approximatifs
  - Vectoriel généralisé
  - Index Sémantic Latent
  - Algorithme Génétique
  - RNA Kohonen
  - RNA Hopfield
- Méthodologie
- Références

# Plan de la présentation

- Concept d'Agent logiciel
- Problématique
- Objectifs du projet
- Modèles de filtrage
- **Méthodologie**
- Références

# Concept d'Agent logiciel

- **Logiciel capable d'actions indépendantes au nom de son utilisateur ou de son propriétaire [Wo02]**
- **Système accomplissant une tâche en interagissant avec son environnement par les moyens de percepteurs et d'effecteurs [Sh94]**
- **Système capable d'acquérir une compétence et de s'adapter aux intérêts changeants de son utilisateur [Ma93]**

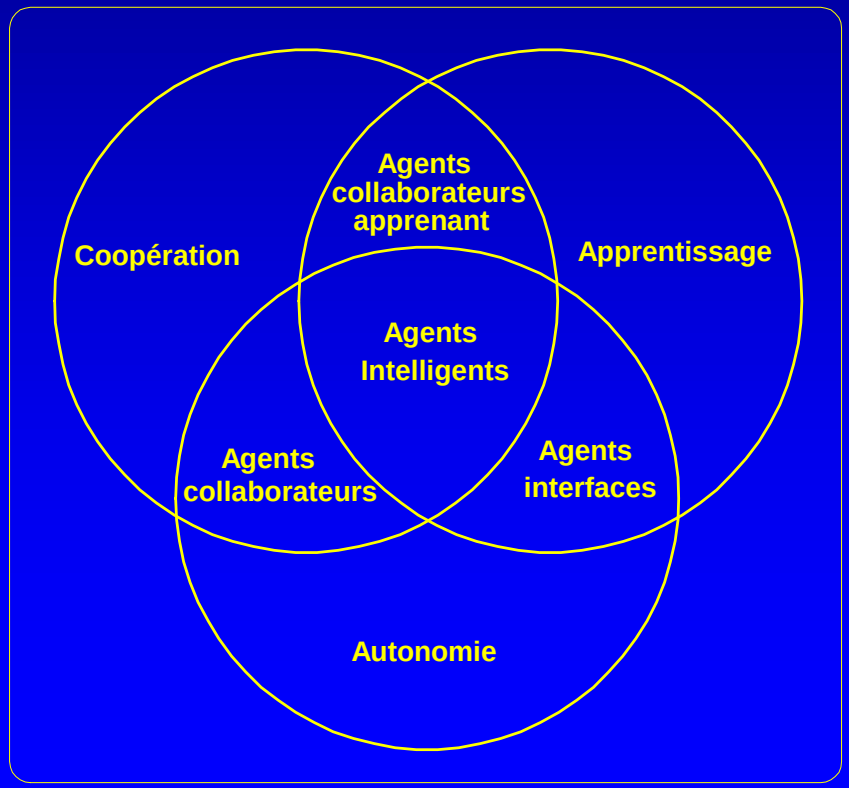
# Concept d'Agent logiciel

- **Un agent autonome est un agent situé dans un environnement dont il fait partie et sur lequel il agit dans le temps, à la poursuite de son propre agenda, de manière à affecter son propre futur [Franklin & Graesser, 1996]**
- **On a autant de chance d'arriver à un consensus sur la définition du concept *"Agent logiciel"* que l'IA a d'arriver à un consensus pour le concept *"Intelligence Artificielle"* [Nw96]**

# Dimensions des Agents logiciels [Nw96]

**3 critères principaux :**

- **Autonomie**
- **Apprentissage**
- **Coopération**




**Autres critères :**

- **Mobilité (statique - mobile)**
- **Réactivité (délibératif - réactif)**
- **Rôle (information, gestion, etc.)**
- **Combinaison (hybrides)**
- **Intérêt (égoïste – altruiste)**
- **Véracité (vérité – mensonge)**

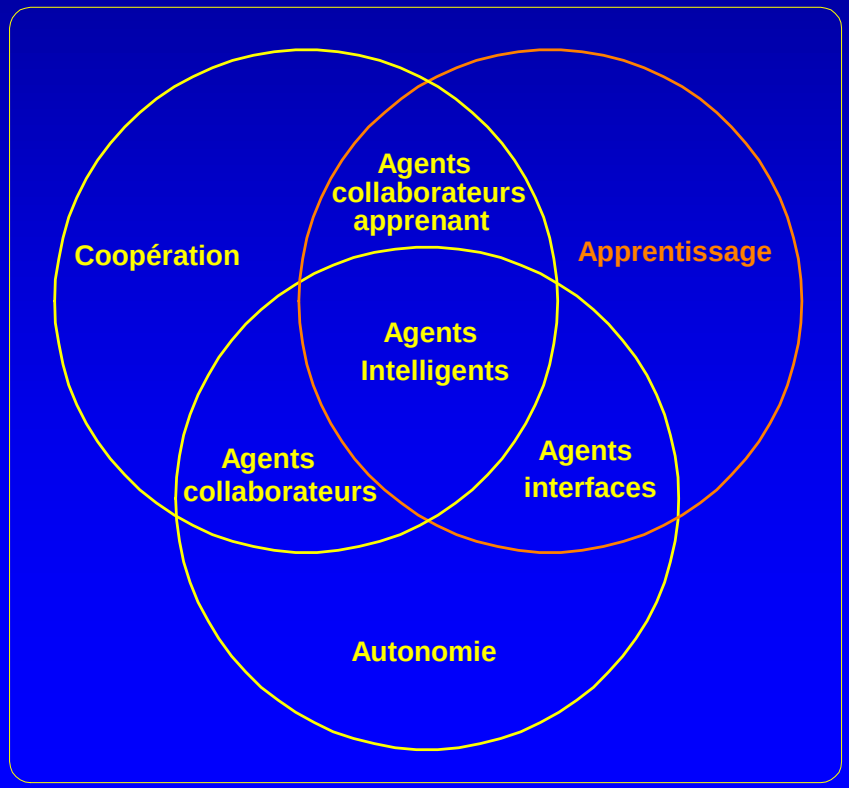
**Architecture générique :**

- **BDI (Belief – Desire – Intention)**
  - **Croyances = connaissances**
  - **Désires = objectifs**
  - **Intentions = actions**
- **Émotion = contrôle des priorités**

# Dimensions des Agents logiciels [Nw96]

**3 critères principaux :**

- **Autonomie**
- **Apprentissage**
- **Coopération**




**Autres critères :**

- **Mobilité (statique - mobile)**
- **Réactivité (délibératif - réactif)**
- **Rôle (information, gestion, etc.)**
- **Combinaison (hybrides)**
- **Intérêt (égoïste – altruiste)**
- **Véracité (vérité – mensonge)**

**Architecture générique :**

- **BDI (Belief – Desire – Intention)**
  - **Croyances = connaissances**
  - **Désires = objectifs**
  - **Intentions = actions**
- **Émotion = contrôle des priorités**

# Typologie des Agents logiciels [Nw96]

**7 agents-types principaux :**

- Collaborateurs : collaboration et autonomie
- Interfaces : apprentissage et autonomie
- Mobiles : téléchargeables
- Information : rôle (filtrage)
- Réactifs : vitesse de réaction
- Hybrides (combinaisons)
- Intelligents : collaboration, apprentissage et autonomie (aucun !)

# Architectures cognitives [Wr94]

- Subsumption (Brooks, 1986) R PP
- Theo (Mitchell *et al*, 1991) R PP
- ATLANTIS (Gat, 1991) R
- MAX (Kuokka, 1991) RD
- AIS (Hayes-Roth *et al*, 1991) RD
- RALPH-MEA (Russell *et al*, 1994) RD
- EPIC (Kieras *et al*, 1997) RD PP
- ACT-R / PM (Anderson *et al*, 1998) RD PP
- LICAI / CoLiDeS (Kitajima & Polson , 1997-2000) RD PP
- Soar (Laird *et al*, 1987) D PP
- ICARUS (Langely *et al*, 1991) D PP
- Teton (VanLehn *et al*, 1991) D PP
- Homer (Vere *et al*, 1990) D
- Prodigy (Carbonell *et al*, 1991) D
- ERE (Drummond *et al*, 1991) D

# Plausibilité Psychologique

"Les architectures cognitives se distinguent des approches de l'ingenierie qui s'efforcent de construire des systèmes intelligents par les techniques qui servent le mieux leurs objectifs. Les architectures cognitives sont conçues pour simuler l'intelligence humaine d'une manière humaine." (Newell, 1990, *Unified Theories of Cognition*).

MHP

# Problématique du filtrage de l'information

- Récupération sélective d'information
  - Indexation des termes
  - Formulation de la requête
  - Appariement requête-documents
- Hypothèse : *texte* = *associations d'informations présentant des régularités identifiables mathématiquement*

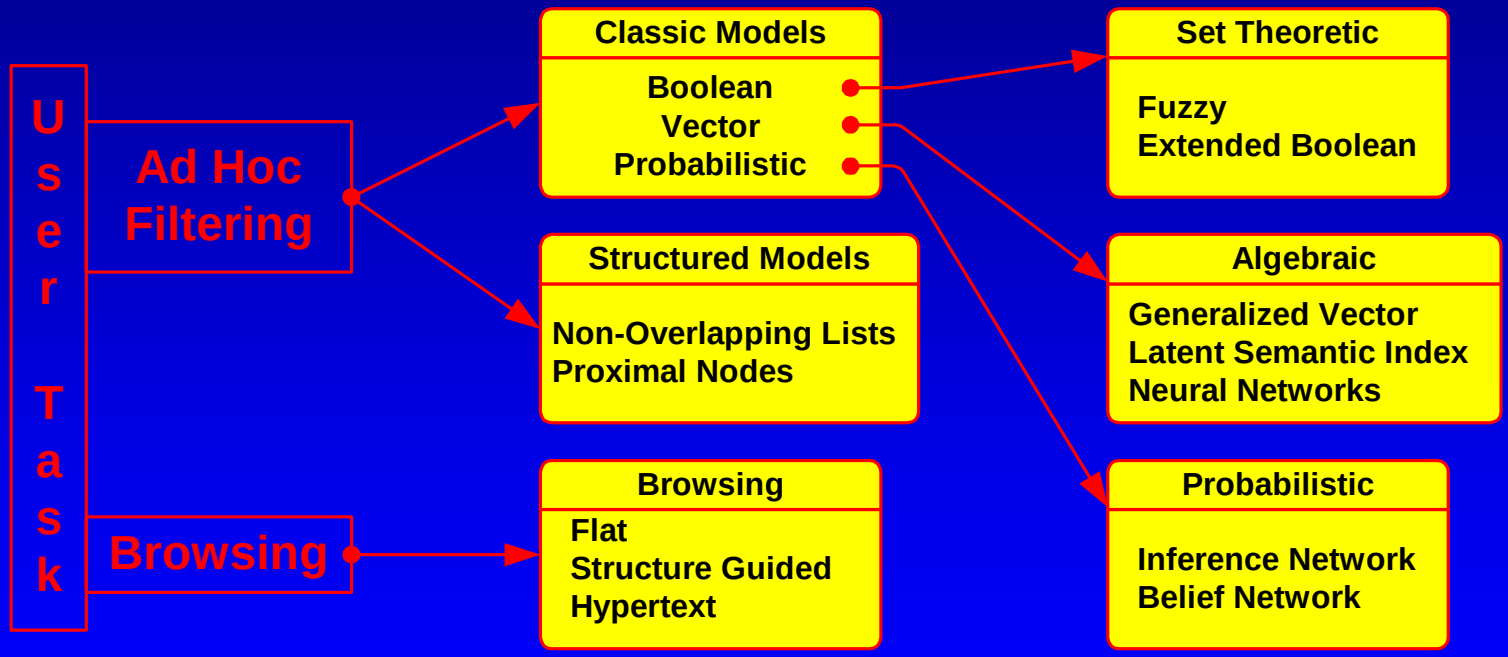
# Problématique du filtrage de l'information

- 32 modèles de filtrage répertoriés
- Éprouvés isolément
- Collections variées pour les essais
- Domaines spécialisés
- Représentations variées
- Quel modèle choisir ?

# Objectifs du projet

- Comparer l'efficacité de filtrage des modèles
  - Revoir tous les modèles
  - Sélectionner les modèles intéressants
  - Identifier un dénominateur commun
  - Déterminer une collection commune
- Évaluer la performance de filtrage
- Évaluer l'extensibilité
- Identifier les forces et faiblesses

# Taxonomie des modèles de filtrage [Ba99]

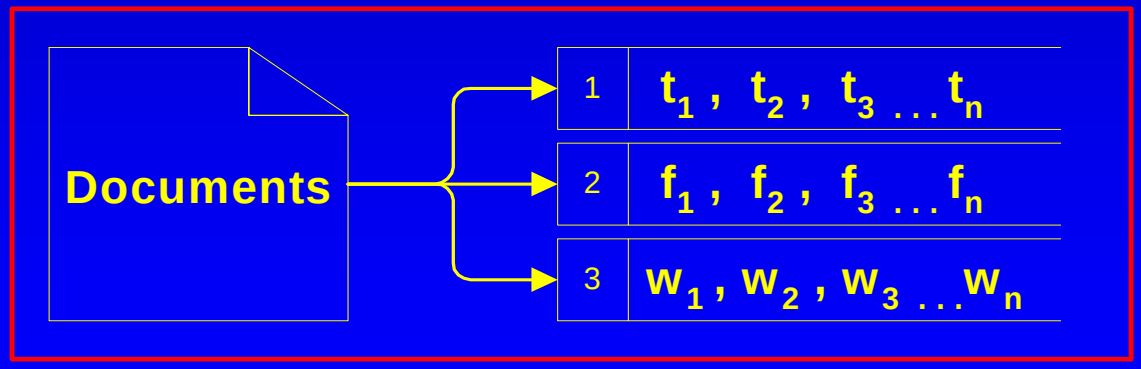


# Modèles de filtrage de l'information

| Conventionnels | Évolutionnaires | Hybrides |
|---|---|---|
| Bayesien naïf | Rétroaction | Géno-probabiliste |
| Baysien classique | Algorithme génétique | Géno-rétroactif |
| Réseau à inférence | Réseaux de neurones | Géno-flou |
| Réseau de croyances | Colonies de fourmis | Neuro-génétique |
| Booléen classique | Système immunitaire | Neuro-flou |
| Booléen flou | Recuit simulé | Neuro-LSI |
| Booléen étendu | Théorie du chaos | Réseau croyances - HITS |
| Ensembliste | | ACO-*K-means* |
| Ensembles approximatifs | | AIS-flou |
| Vectoriel classique | | Multi-Hopfield |
| Vectoriel généralisé | | |
| Index sémantique latent | | |

# Modèles conventionnels
# 1. Vectoriel classique [Sa71]

- Requêtes et documents
  - document = { termes }
  - $tf_i$ : fréquence du $i^{\text{ième}}$ terme
  - $idf_i$ : fréquence documentaire inverse du $i^{\text{ième}}$ terme
  - $w_{ik}$ : poids du $i^{\text{ième}}$ terme dans le document $k$



# Modèles conventionnels
## 1. Vectoriel classique [Sa71]

- Pondération des termes par document

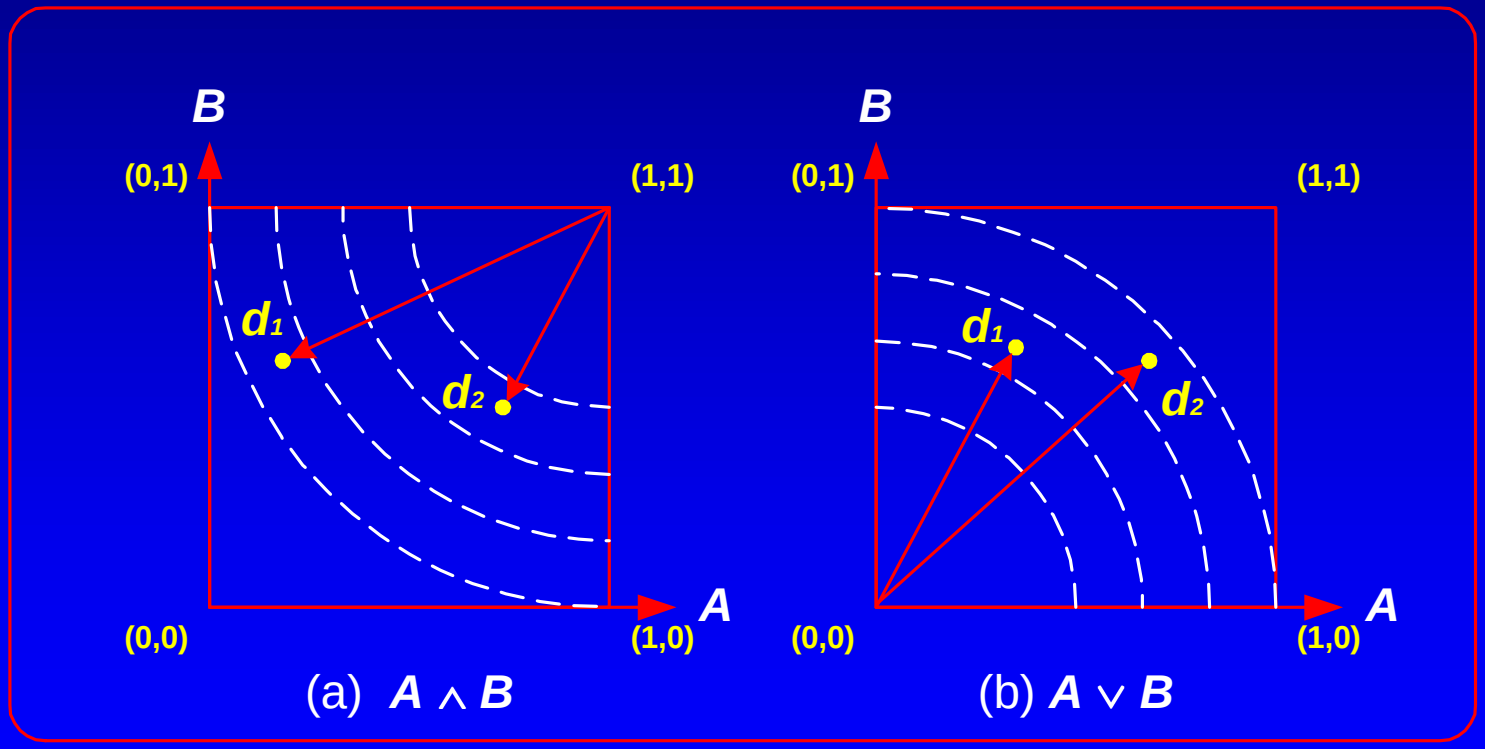
$$w_{i,k} = tf_{i,k} \times idf_i = \frac{tf_{i,k}}{\max_k tf_{i,k}} \times \log \frac{N}{n_i}$$

- Appariement requête-documents

$$sim(q, d_k) = \frac{\sum_{i=1}^{n} w_{i,k} \times w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,k}^2} \times \sqrt{\sum_{i=1}^{n} w_{i,q}^2}}$$

# Modèles conventionnels
# 2. Booléen Étendu [Sa83a]

- Principe



# Modèles conventionnels
# 2. Booléen Étendu [Sa83a]

- Booléen + vectoriel (pondérations, similarité)
- Poids $w_t = tf_t \times idf_t$
- Similarité = distance selon théorie ***p-norm***

$$q_\vee = t_1 \vee^p t_2 \vee^p \ldots \vee^p t_m \quad sim(q_\vee, d) = \left( \frac{w_{t1}^{\,p} + w_{t2}^{\,p} + \ldots + w_{tm}^{\,p}}{m} \right)^{\frac{1}{p}}$$

$$q_\wedge = t_1 \wedge^p t_2 \wedge^p \ldots \wedge^p t_m \quad sim(q_\wedge, d) = 1 - \left( \frac{\left(1 - w_{t1}^{\,p}\right) + \left(1 - w_{t2}^{\,p}\right) + \ldots + \left(1 - w_{tm}^{\,p}\right)}{m} \right)^{\frac{1}{p}}$$

- Combinaisons possibles : $\left(t_1 \vee^2 t_2\right) \wedge^\infty t_3$

# Modèles conventionnels
## 3. Ensembliste [Po02]

- Extraction de sémantique dans la cooccurrence des termes
- Algorithme *Apriori* [Mi97] : ensembles fréquents de termes
- Documents indexés par ces ensembles fréquents = concepts

$$w_{i,j} = sf_{i,j} \times ids_i = sf_{i,j} \times \log \frac{N}{ds_i}$$

- Similarité par une mesure du cosinus entre les vecteurs d'ensembles fréquents

# Modèles conventionnels
# 4. Ensembles approximatifs [Pa82; Da88]

- Théorie des ensembles approximatifs [Pa82] :
  - limite supérieure $\overline{A}(o)$: plus petit $\subset A$ qui contient $o$
  - limite inférieure $\underline{A}(o)$ : plus grand $\subset A$ qui contient $o$
- Segmentation des termes du vocabulaire en concepts $\{C_1, C_2, \ldots C_n\}$
  - Utilisation de l'algorithme *Apriori*
- Document $D_i$ représenté par les termes $X_i$
- $\underline{A}(X_i) \subseteq A(X_i) \subseteq \overline{A}(X_i)$

# Modèles conventionnels
# 4. Ensembles approximatifs [Pa82; Da88]

- Variété de stratégies de filtrage
  - $\underline{A}(Q) = \underline{A}(X_i)$
  - $\overline{A}(Q) = \overline{A}(X_i)$
  - $\overline{A}(Q) = \underline{A}(X_i)$
  - $\underline{A}(Q) = \overline{A}(X_i)$
  - $\underline{A}(Q) \subseteq \underline{A}(X_i)$
- Similarité calculée par
  - $\mathrm{SIM}(Q, D) = \underline{\mathrm{SIM}}(Q, D) + \overline{\overline{\mathrm{SIM}}}(Q, D)$
  - $\underline{\mathrm{SIM}}(Q, D) = |\underline{A}(Q) \cap \underline{A}(X)| \;/\; |\underline{A}(Q) \cup \underline{A}(X)|$
  - $\overline{\overline{\mathrm{SIM}}}(Q, D) = |\overline{A}(Q) \cap \overline{A}(X)| \;/\; |\overline{A}(Q) \cup \overline{A}(X)|$

# Modèles conventionnels
# 5. Vectoriel Généralisé [Wo85]

- Indépendance des termes indexés deux à deux n'implique pas qu'ils soient orthogonaux
- Chaque document est représenté par un vecteur $m_i$ (*minterm*) décomposables en vecteurs orthogonaux $\vec{m}_i$

$$m_1 = (0,0,...,0)$$
$$m_2 = (1,0,...,0)$$
$$\vdots$$
$$m_{2^t} = (1,1,...,1)$$

- $m_1$ pointe aux documents ne contenant aucun des termes indexés
- $m_2$ pointe aux documents contenant seulement le premier terme indexé
- $m_{2^t}$ pointe aux documents contenant tous les termes indexés

# Modèles conventionnels
## 5. Vectoriel Généralisé [Wo85]

- Vecteur $\vec{t}_i$ associé au terme $t_i$ est la somme normalisée des vecteurs $\vec{m}_r$ activés pour ce $t_i$
- $c_{i,r}$ est le facteur de corrélation du terme $t_i$ associé aux documents pointés par $m_r$
- Le produit interne $\vec{t}_i \bullet \vec{t}_j$ donne le degré de corrélation entre deux termes $t_i$ et $t_j$

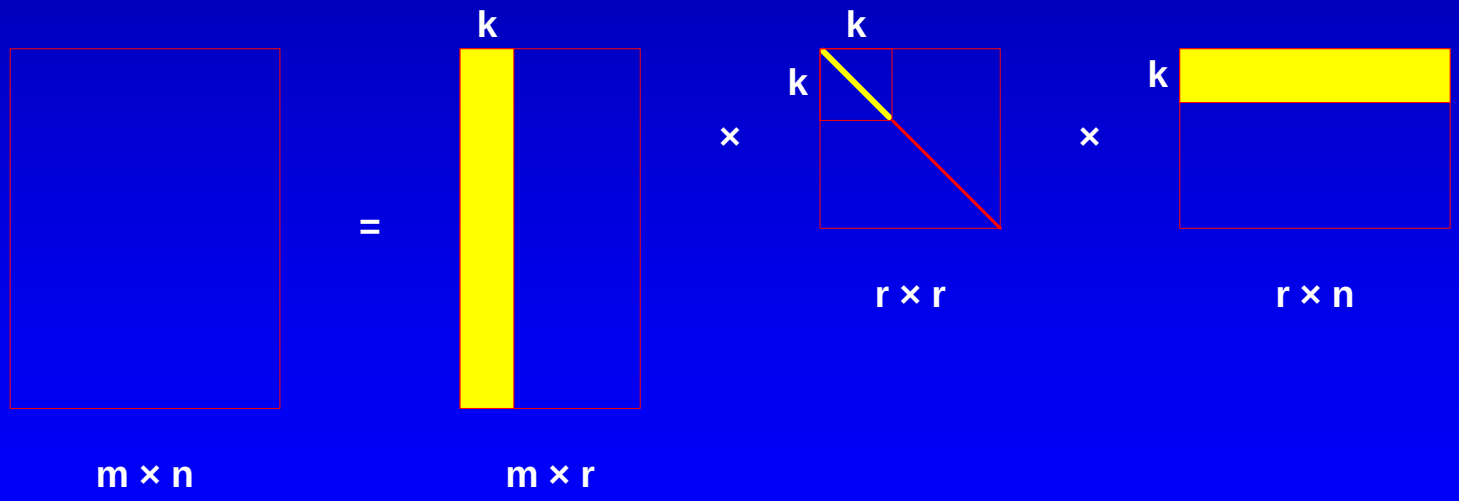
$$\vec{t}_i = \frac{\sum_{\forall r, g_i(m_r)=1} c_{i,r}\vec{m}_r}{\sqrt{\sum_{\forall r, g_i(m_r)=1} c_{i,r}^2}}$$

$$c_{i,r} = \sum_{d_j | \forall l, g_l(\vec{d}_j)=g_l(m_r)} w_{i,j}$$

$$\vec{t}_i \bullet \vec{t}_j = \sum_{\forall r | g_i(m_r)=1 \wedge g_j(m_r)=1} c_{i,r} \times c_{j,r}$$

# Modèles conventionnels
# 6. Index Sémantique Latent [De90; Be95; Pa97]

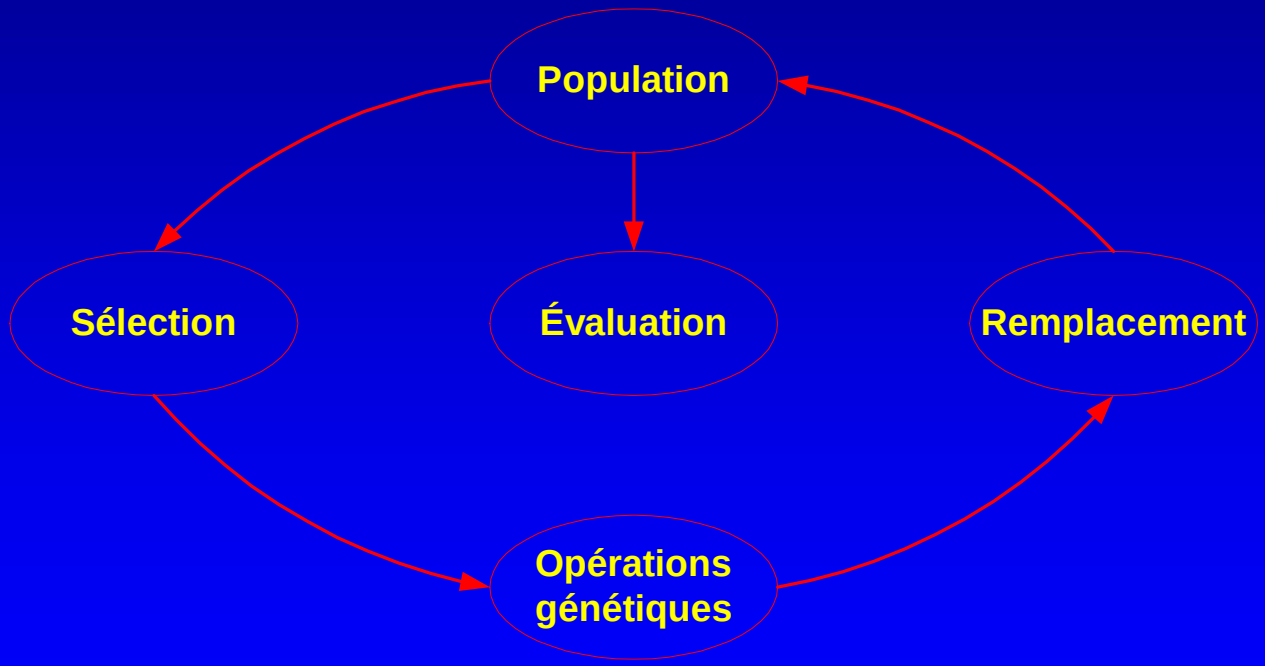
- **Matrice termes-documents : $A_{m \times n}$**
- **Compression à k dimensions : $A_k = U_k D_k V_k^T$ (SVD)**
- **Projection de la requête : $q^* = q^T U_k D_k^{-1}$**




$$A_{m \times n} = U_{m \times r} \times D_{r \times r} \times V_{n \times r}^T$$

# Modèles évolutionnaires
# 7. Algorithme Génétique [Ho75; Go88; De00]

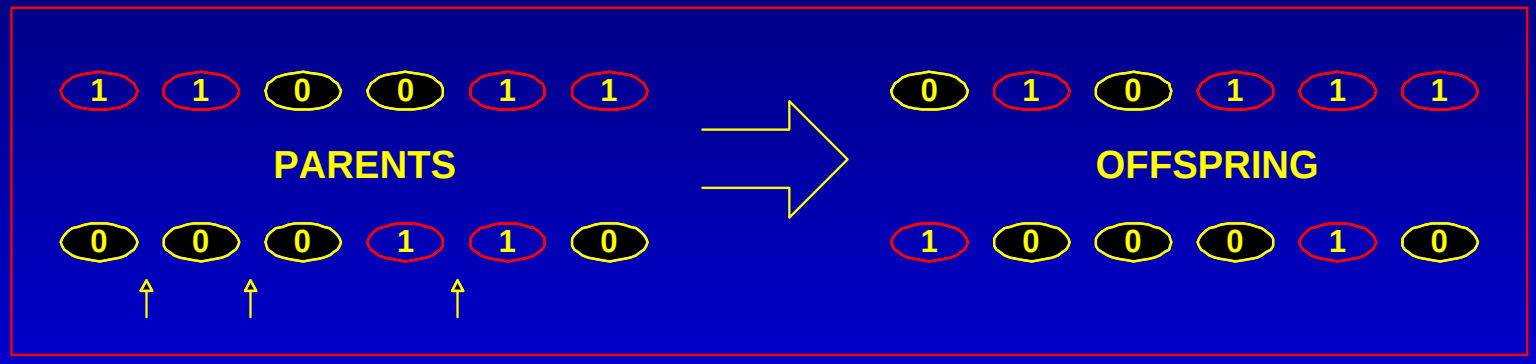
- Cycle génétique



# Modèles évolutionnaires
## 7. Algorithme Génétique [Ho75; Go88; De00]

- Population ≡ profil d'intérêt (requêtes)
- Individu ≡ vecteur de termes de la requête
- Gène ≡ terme (poids)
- Sélection naturelle ≡ fonction objective (fonctions de similarité \ conformité)

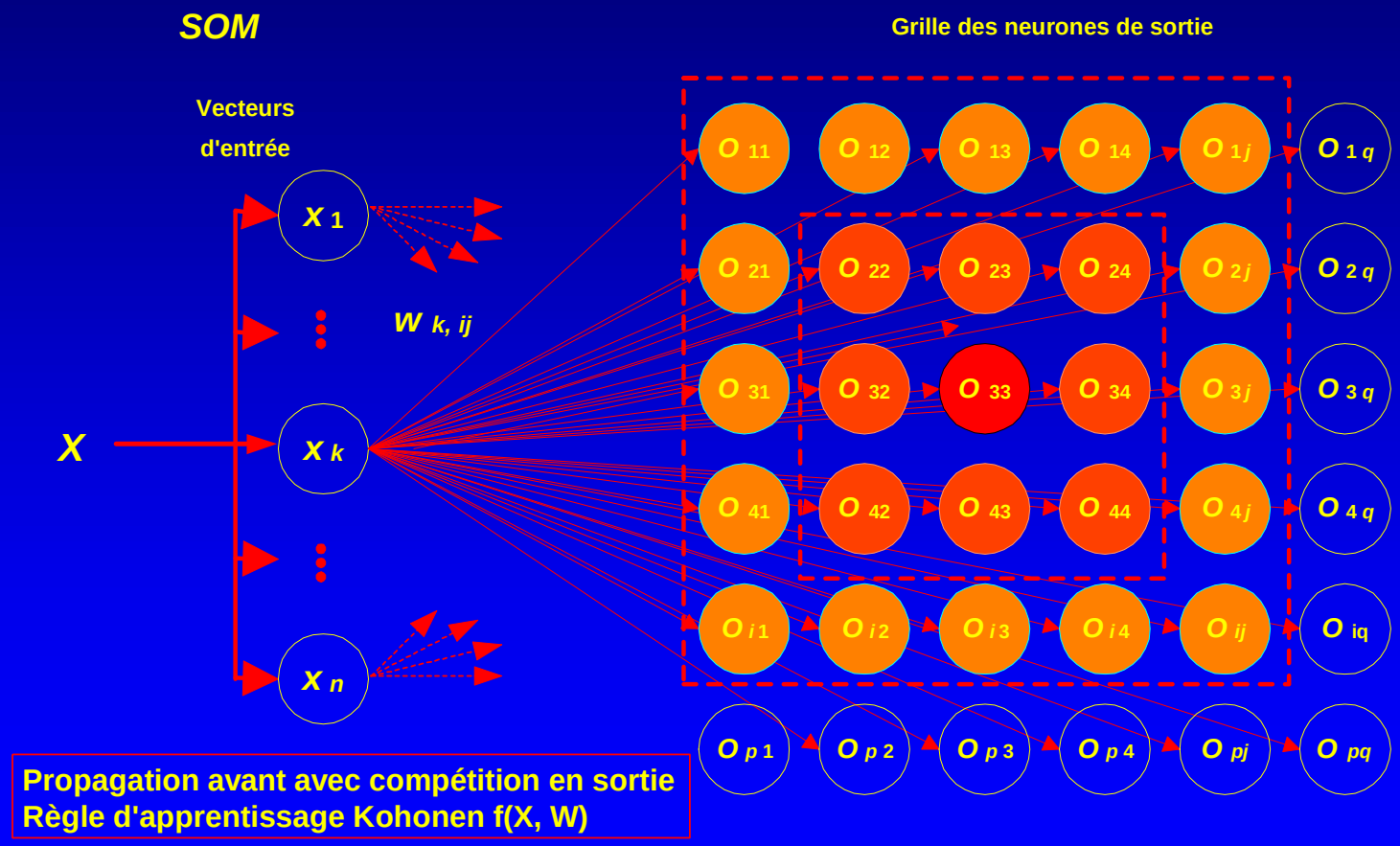
# Modèles évolutionnaires
# 7. Algorithme Génétique [Ho75; Go88; De00]




- **Croisement :** interchange d'une partie du code génétique
- **Mutation :** modification d'un gène aléatoire

# Modèles évolutionnaires

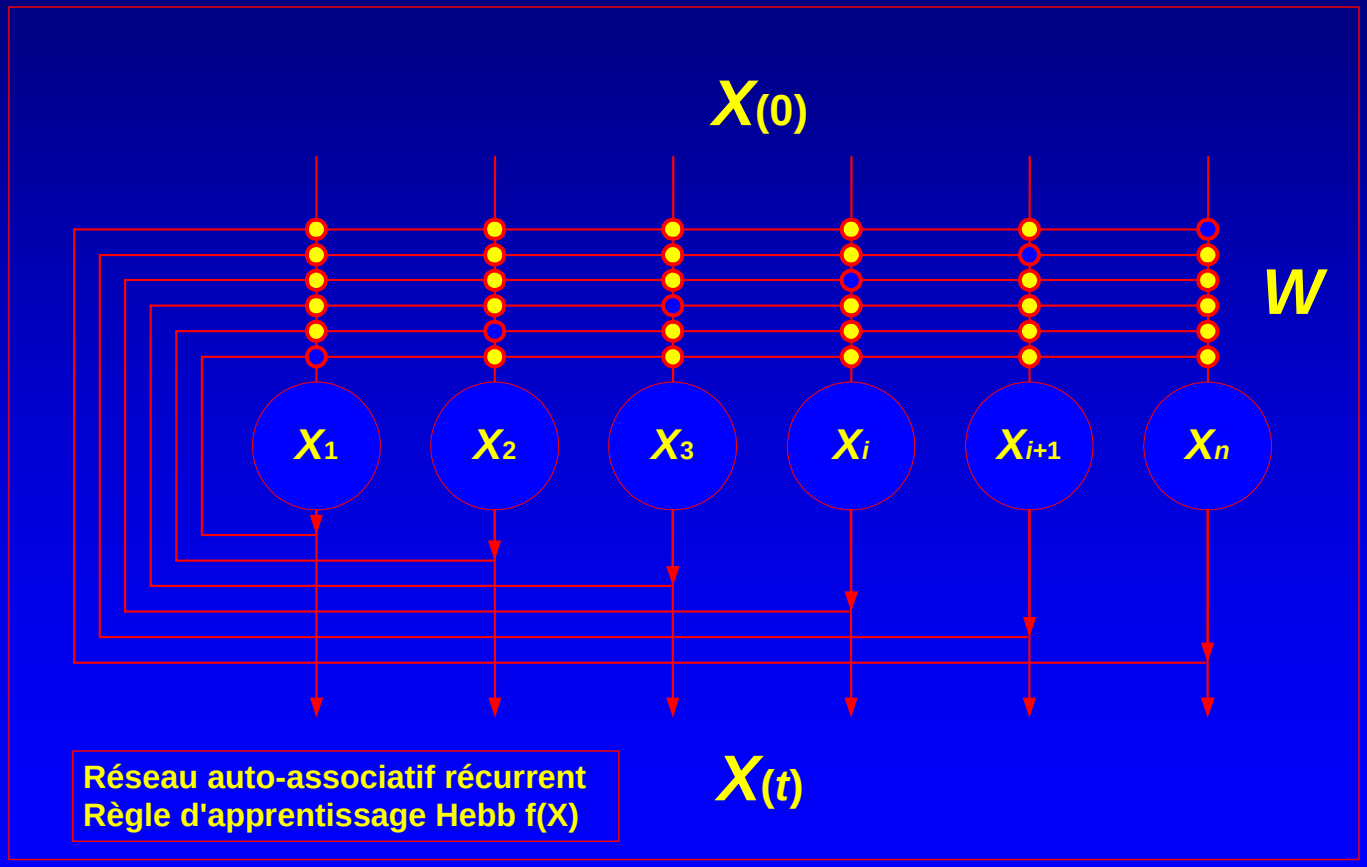
## 8. RNA de Kohonen [Kohonen 82; Li91; Kw95]




**Propagation avant avec compétition en sortie**
**Règle d'apprentissage Kohonen f(X, W)**

# Modèles évolutionnaires

## 9. RNA de Hopfield [Hopfield 82; Ch94b]



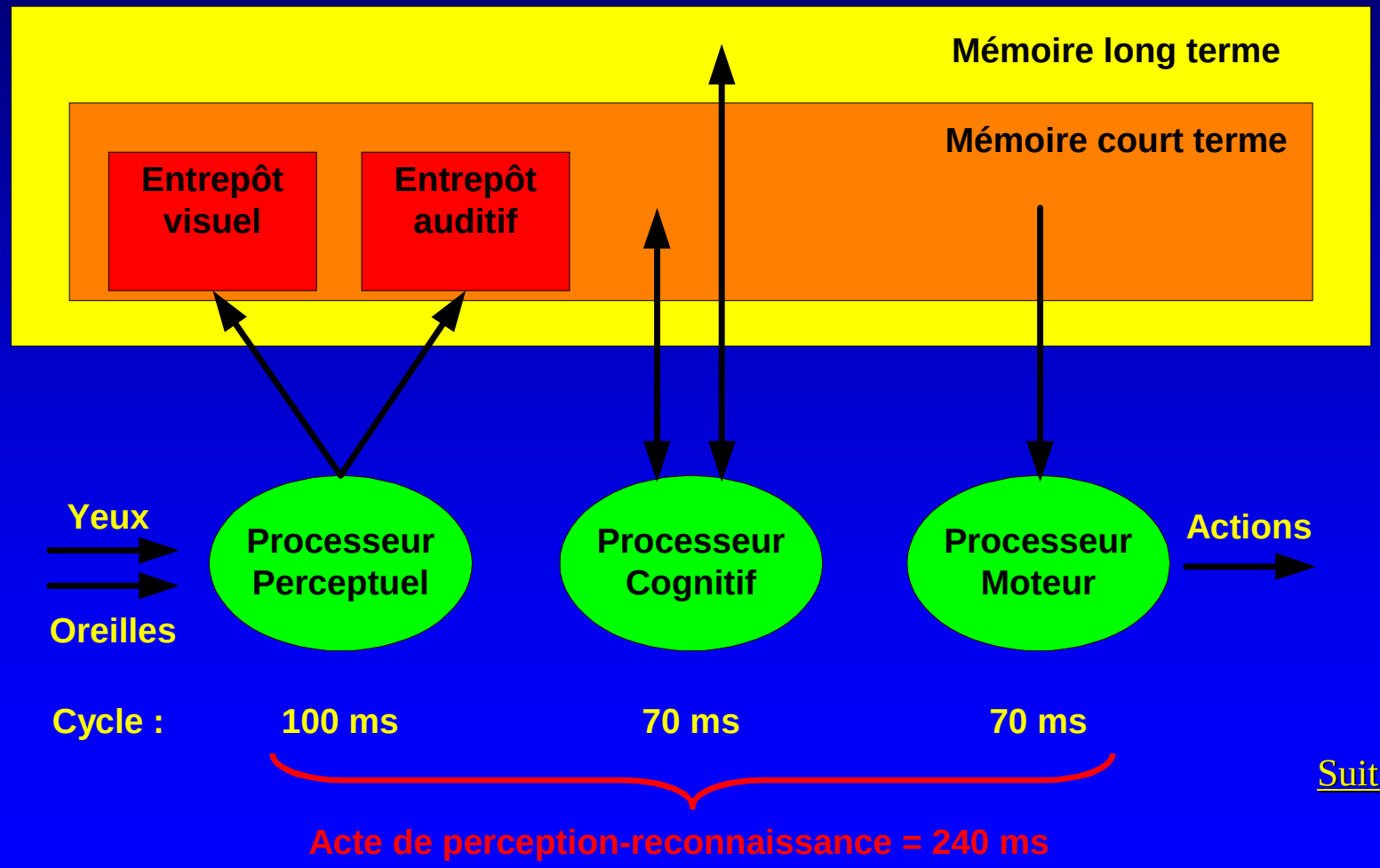
# Méthodologie

- TREC ("*Text REtrieval Collection*")
  - 2 millions de documents catégorisés
  - 500 sujets de requête
- Mesures de rappel – précision classiques
- Courbes rappel-précision
- Moyenne harmonique ("*F-score*")
- Environnement : IntellAgent [De00]

# Références

Microsoft Word Document

# Modèle du Processeur Humain [Card, 83]




Suite