ACI MASSES DE DONNEES - PROJET MD33/04-07 - APMD: ACCES PERSONNALISE A DES MASSES DE DONNEES

http://apmd.prism.uvsq.fr

ACCÈS PERSONNALISÉ À DES MASSES DE DONNÉES

2004 - 2007

PARTENAIRES DU PROJET

CLIPS-IMAG, Grenoble
- Catherine Berrut
- Nathalie Denos
- An Te Nguyen

IRISA, Lannion
- Patrick Bosc
- Daniel Rocacher
- Ludovic Liétard

IRIT, Toulouse
- Mohand Boughanem
- Chantal Soulé-Dupuy
- Lynda Lechani
- Pascaline Tchienehom
- Nesrine Zemirli

LINA, Nantes
- Noureddine Mouaddib
- Guillaume Raschia
- Laurent Ughetto
- Amenel Voglozin

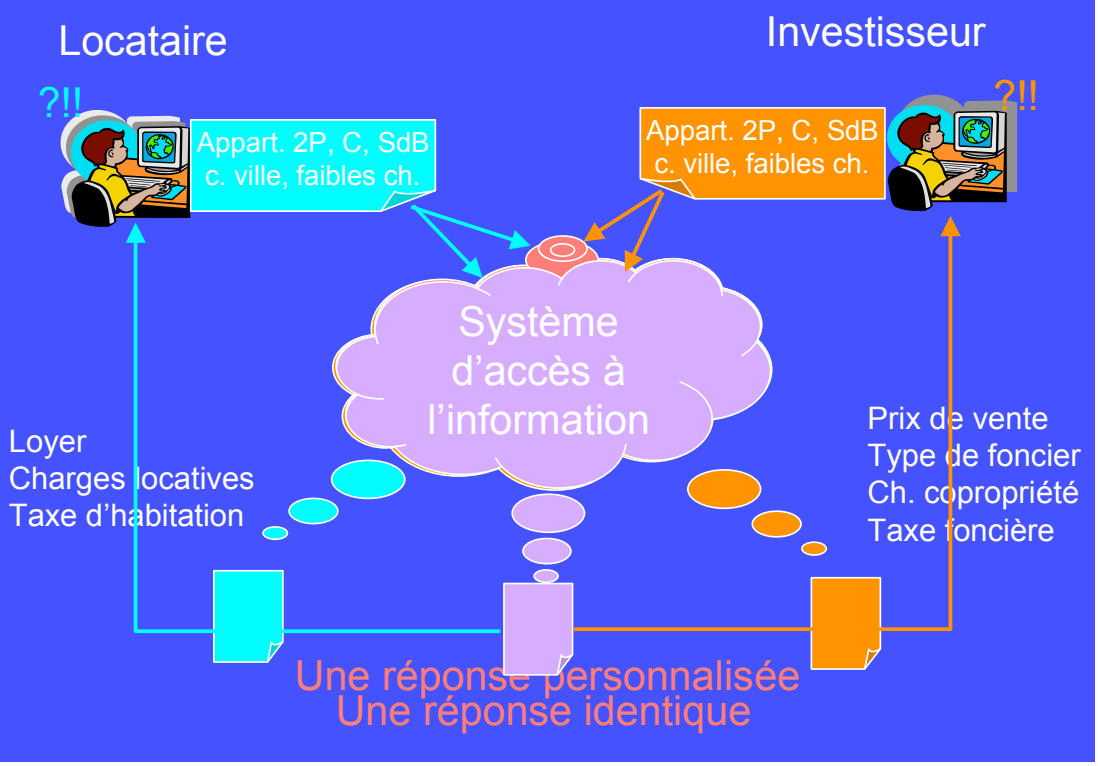
LIRIS, Lyon
- Sylvie Calabretto
- Béatrice Rumpler
- Hassan Nadery
- Souela Bohé

PRiSM, Versailles
- Mokrane Bouzeghoub
- Stéphane Lopes
- Dimitre Kostadinov
- Veronika Peralta

PROBLÉMATIQUE

- **Prolifération des sources de données distribuées**
  - résultats massifs en réponse aux requêtes des utilisateurs (surcharge d'informations)
  - temps de réponse de plus en plus longs (passage à l'échelle difficile)
- **Modèles d'accès uniformes : les systèmes ne tiennent pas suffisamment compte**
  - du contexte et des préférences de l'utilisateur
  - des caractéristiques des sources de données (qualité)

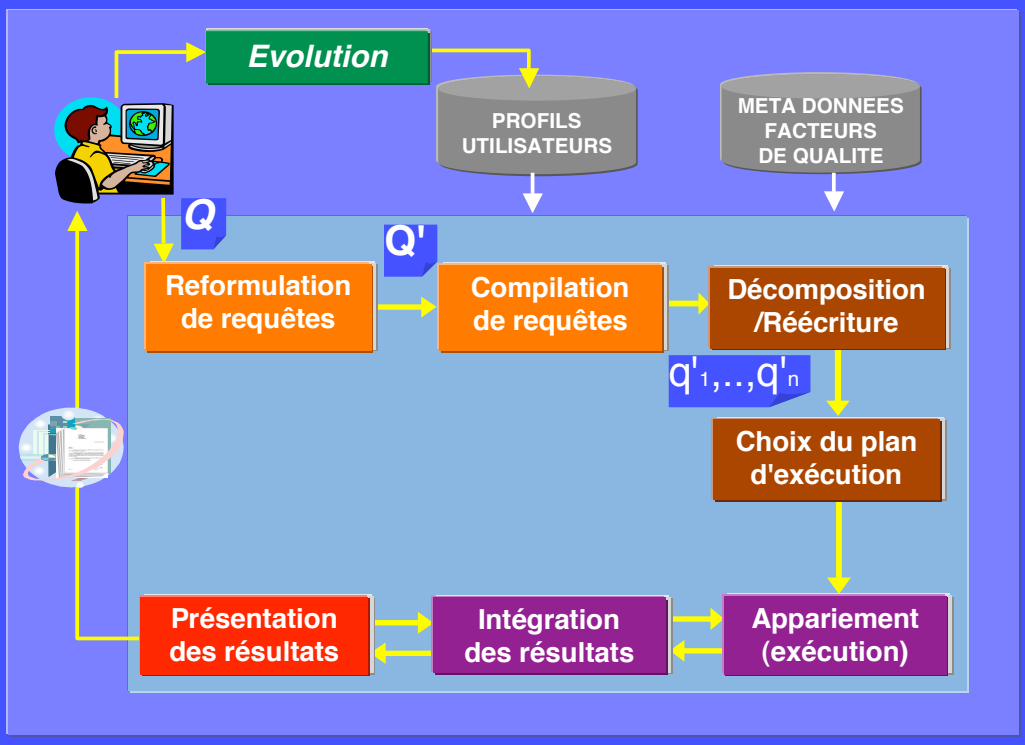
MODÈLES D'ACCÈS UNIFORMES



OBJECTIFS

- **Placer la personnalisation de l'information au centre des SRI et des SGBD afin de**
  - simplifier la formulation des requêtes
  - améliorer la pertinence des informations délivrées par les systèmes
  - maîtriser le passage à l'échelle
- **Exploiter les profils des utilisateurs et la qualité des sources d'information**
  - centres d'intérêt, préférences, environnement des utilisateurs
  - fraîcheur, disponibilité, crédibilité des sources d'information
- **Agir sur l'ensemble du cycle de vie des processus d'accès à l'information**

AGIR SUR TOUT LE CYCLE DE VIE D'UNE REQUÊTE



APPROCHE DU PROJET

- **Exploration systématique et aussi large que possible de la notion de profil**
  - de quoi est constitué un profil, comment le construire, comment l'utiliser, comment le faire évoluer
- **Identification des facteurs de qualité influant sur la personnalisation**
  - quels sont les facteurs de qualité liés à l'information recherchée, aux sources fournissant cette information, aux processus de production de cette information, aux modes d'exécution des requêtes, au contexte d'interrogation de l'utilisateur
- **Définition d'un modèle adaptatif d'exécution de requêtes**
  - quelle partie du profil influe sur quelle étape du cycle de vie d'une requête, comment est évaluée la qualité lors de l'exécution, comment intégrer le feedback de l'utilisateur, quelles traces garder pour justifier les réponses du système, quelle influence ont le profil et la qualité sur la réduction de l'espace de recherche

ACTIVITÉ DE LA PREMIÈRE ANNÉE

- **Effort de convergence de communautés différentes sur une problématique commune**
  - États de l'art (pas triviaux sur un sujet vaste et éclaté)
  - Typologie commune des connaissances de profils
  - Modèles d'expression de préférences
  - Typologie des facteurs de qualité pris en compte dans la personnalisation
- **Quelques prototypes et expérimentations réalisés ou en cours**
  - Plateforme de définition et de gestion de profils
  - Génération de profils à partir de résumés de programmes TV
  - Interaction via des cartes de communautés
  - Accès personnalisé à des bibliothèques numériques (thèses)
  - Modèle «Push» personnalisé

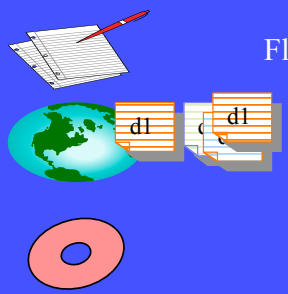
Modèle de «Push» personnalisé

M. Boughanem

Contexte : «Push» = SFI

Sources

Utilisateurs ayant des besoins définis au préalable = Profils



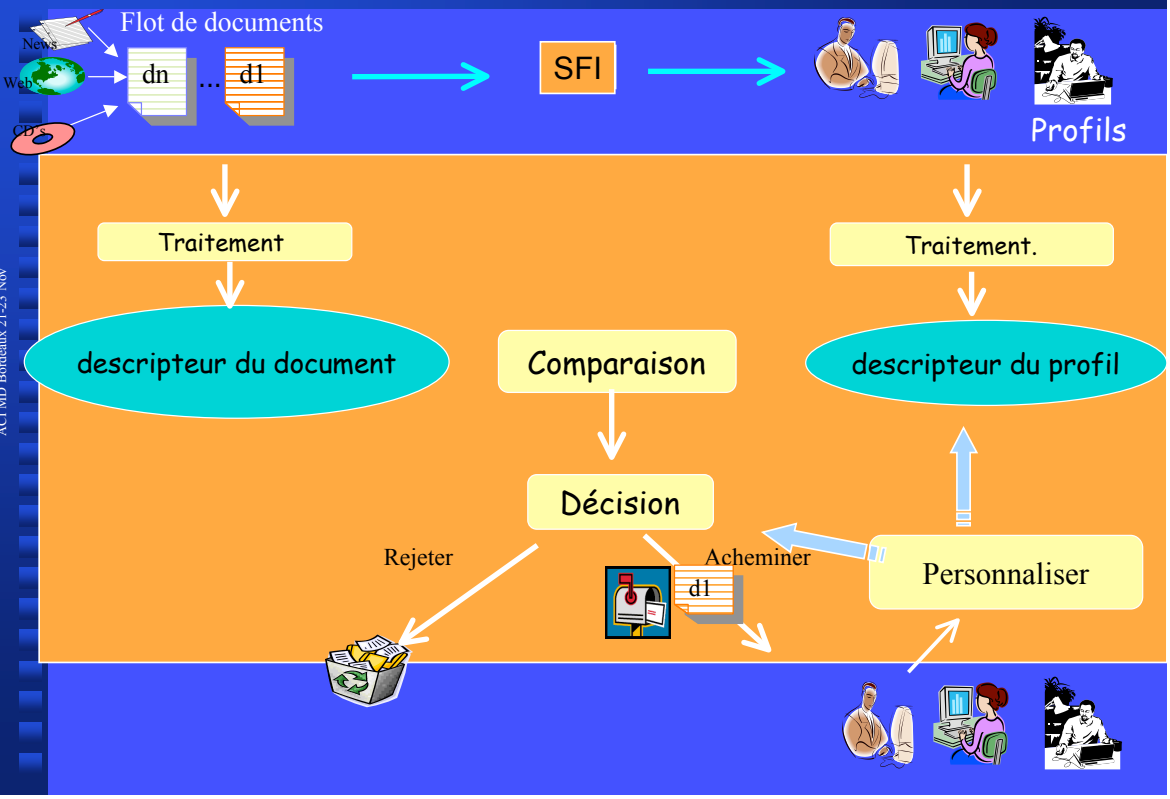
Flot de documents

SFI/push

Terminologie :

- Push
- Filtrage d'information
- Recommandation
- Dissémination sélective d'information
- Système d'alerte

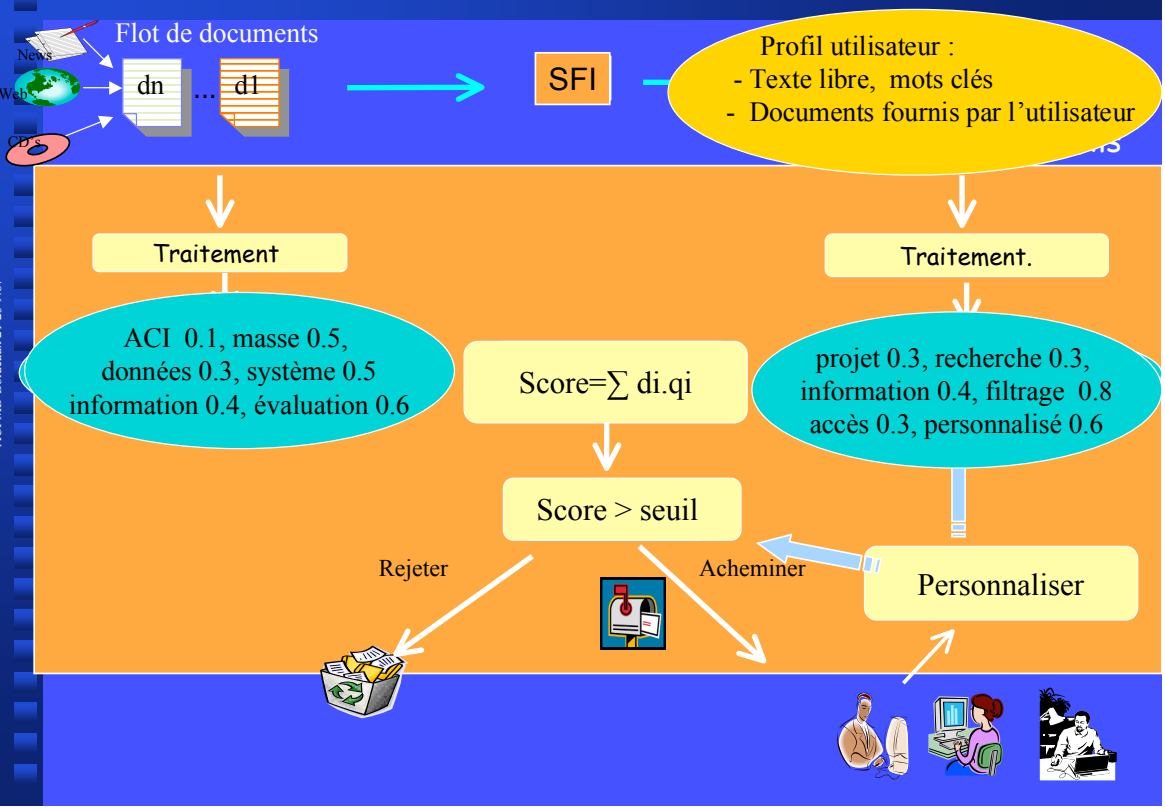
Filtrage d'information = Push



Plan

- Contexte du travail
- **Personnaliser le processus de FI**
  - Adaptation du profil
  - Adaptation de la fonction de décision
- **Expérimentations et résultats**
- **Conclusion**

Filtrage d'information textuelle



Récap. filtrage d'info. textuelle

- **Document :** $p^{(t)} = \lfloor (t_1, w_1^{(t)}), \cdots, (t_n, w_n^{(t)}) \rfloor$
- **Profil :** $d^{(t)} = \lfloor (t_1, d_1^{(t)}), \cdots, (t_n, d_n^{(t)}) \rfloor$
- **Score :** $rsv\left(d^{(t)}, p^{(t)}\right) = \sum_{t_i \in d^{(t)}, tp_j \in p, t_i = tp_j} d_i^{(t)} * w_j^{(t)}$
- **Décision :** $\begin{cases} \text{si } rsv\left(d^{(t)}, p^{(t)}\right) \geq seuil^{(t)} \text{ sélectionner } d^{(t)} \\ \text{sinon rejeter } d^{(t)} \end{cases}$

Personnaliser le processus de FI

- **Processus effectué de manière incrémentale**
  - Dès qu'un document est jugé pertinent par l'utilisateur
  - Faire évoluer le profil
  - Faire évoluer la fonction de décision (construire un nouveau seuil)

Adaptation profil (1/4)

- **Ajouter et/ou supprimer des termes et/ou ajuster les poids des termes dans la représentation du profil**
- **Idée de base**
  - Trouver le profil « temporaire $p_x^{(t)}$ » qui permet de sélectionner le document pertinent, $d_i^{(t)}$, avec un score élevé

$$rsv(d^{(t)}, p_x^{(t)}) = \sum_{i=1}^{n} d_i^{(t)} * p_x^{(t)} = \lambda$$

  - Intégrer le profil temporaire dans le profil global

$$p^{(t+1)} = f(p^{(t)}, p_x^{(t)})$$

Adaptation profil (2/4)

- Définitions
  - Profil idéal : **profil qui permet de sélectionner que des documents pertinents**
  - Poids idéal : **poids d'un terme dans le profil idéal, soit** $f^{(t)}(t_i)$
- Contrainte :
  - **« Les termes du profil temporaire, solution de l'équation, doivent contribuer de manière proportionnelle à leur importance réelle dans le profil idéal.**

Traduction Formelle:

$$\frac{p_x w_i^{(t)}}{f^{(t)}(t_i)} = cste$$

ACI MASSES DE DONNEES - PROJET MD33/04-07 - APMD: ACCES PERSONNALISE A DES MASSES DE DONNEE
ACI MD Bordeaux 21-23 Nov

Adaptation profil (3/4)

1. Réécriture de l'équation

$$\begin{cases} \sum_{i=1}^{n} d_i^{(t)} * p_x w_i^{(t)} = \lambda \\ \forall (t_i, t_j) \in D_j^{(t)} \times D_j^{(t)} \,/\, \dfrac{p_x w_i^{(t)}}{f^{(t)}(t_i)} = \dfrac{p_x w_j^{(t)}}{f^{(t)}(t_j)} \end{cases}$$

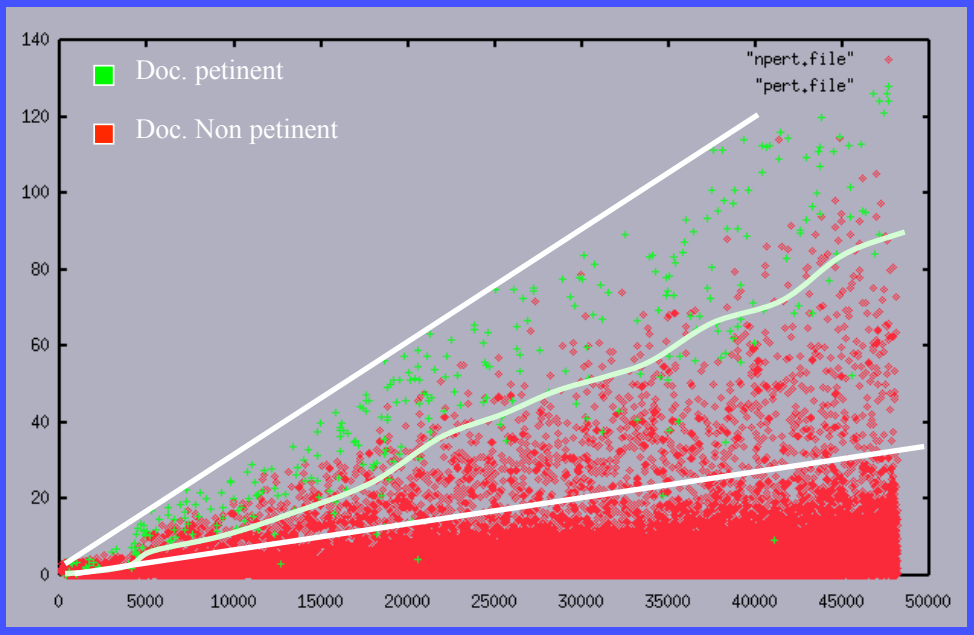
$$p_x w_i^{(t)} = \frac{\lambda . f_i^{(t)}}{\sum_{k=1}^{n} f_k^{(t)} . d_k^{(t)}}$$

2. Intégrer le profil temporaire dans le profil global

$$\forall i \,/\, w_i^{(t+1)} = h(w_i^{(t)}, p_x w_i^{(t)}) = w_i^{(t)} + \alpha * \log(1 + p_x w_i^{(t)})$$

Adaptation du profil (4/4)

Évolution des scores entre un profil donné et des documents



Adaptation de la fonction de décision (1/8)

- Adaptation du processus de décision (seuil) à chaque arrivée de documents pertinents
- Idée de base
  - **Trouver le seuil qui sélectionne le maximum de documents pertinents et le minimum de documents non pertinents**
  - **Fonction d'utilité :**

$$U_{(\lambda_1,\lambda_2,\lambda_3,\lambda_4)}(\theta) = \lambda_1 R_+(\theta) + \lambda_2 S_+(\theta) + \lambda_3 R_-(\theta) + \lambda_4 S_-(\theta)$$

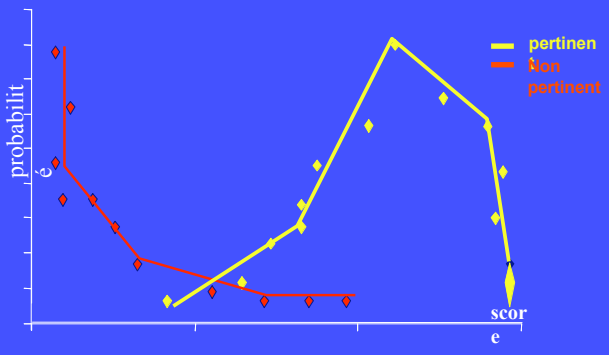
- Exemples

$$U = R_+ - S_+ \qquad T11U = 2.R_+ - S_+$$

Adaptation de la fonction de décision (2/8)

1. Construire (ou dessiner) les distributions de probabilités des scores des documents pertinents et non pertinents ➔ régression linéaire

   $P_r(x)$ et $P_{nr}(x)$



Adaptation de la fonction de décision (3/8)

1. Construire (ou dessiner) les distributions de probabilités des scores des documents pertinents et non pertinents ➔ régression linéaire

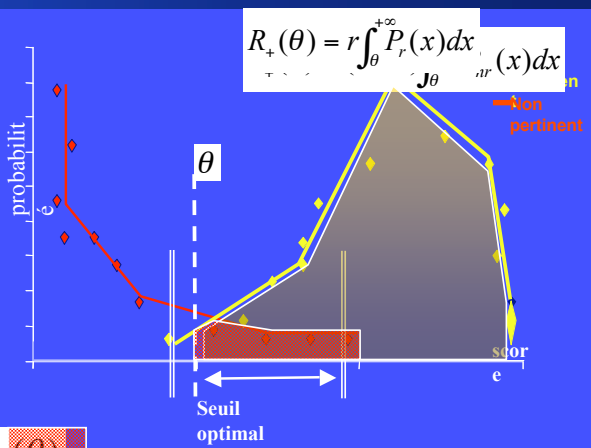
   $P_r(x)$ et $P_{nr}(x)$

2. Réécrire la fonction d'utilité

$$U_{(\lambda_1,\lambda_2)}(\theta) = \lambda_1 R_+(\theta) + \lambda_2 S_+(\theta)$$

3. Déterminer le seuil « idéal » par maximisation de la fonction d'utilité

$$\theta^* = \arg\max_{\theta} U_{(\lambda_1,\lambda_2)}(\theta)$$



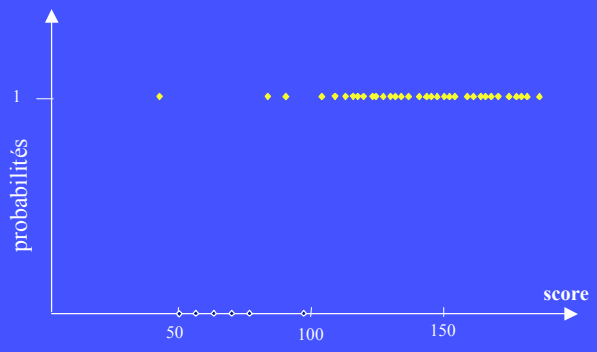
Adaptation de la fonction de décision (4/8)

1. Construction de distributions de probabilités

- Conversion des scores en probabilités

$$\forall score_x \in [score_{\min}, score_{\max}] /$$

$$p(X = score_x) = \frac{\left| \left\{ D_j^{(t)} / rsv(D_j^{(t)}, P^{(t)}) = score_x \right\} \right|}{\left| \left\{ E_r^{(t)} \right\} \right|}$$

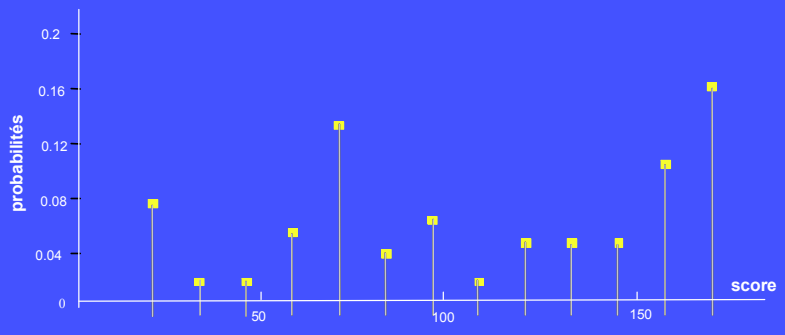


Adaptation de la fonction de décision (5/8)

1. Construction de distributions de probabilités

- Conversion des scores en probabilités
  - Estimation des probabilités par intervalle
  - Intervalles d'amplitudes égales établies à partir de l'écart type

$$p(score_{i-1} \le X < score_i) = \frac{\left| \left\{ D_j^{(t)} / rsv(D_j^{(t)}, P^{(t)}) \in I_i \right\} \right|}{\left| \left\{ E_r^{(t)} \right\} \right|}$$

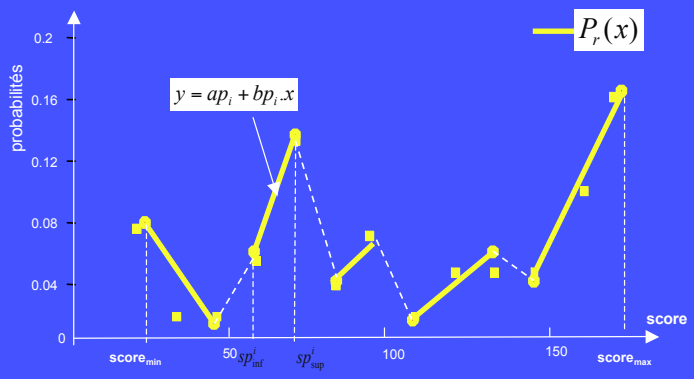


Adaptation de la fonction de décision (6/8)

1. Construction de distributions de probabilités

- Conversion des scores en probabilités
- Estimation des probabilités par intervalle
- Linéarisation de la distribution des scores

• *Chaque segment de droite :* $C_p^i(sp_{\inf}^i, sp_{\sup}^i, ap_i, bp_i)$



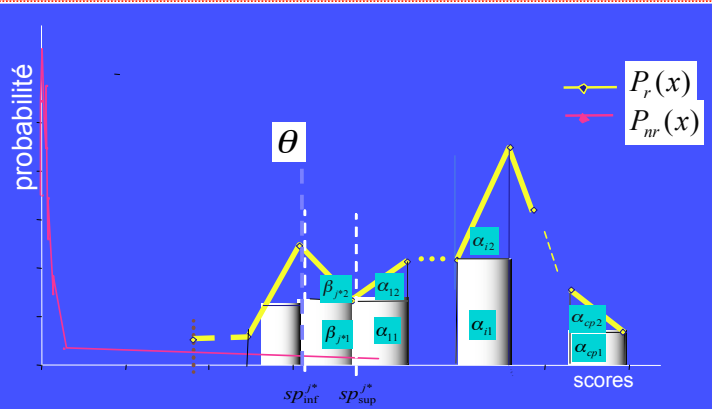
Adaptation de la fonction de décision (7/8)

2. Réécrire la fonction d'utilité

$$\theta^* = \arg\max_{\theta}\left[(\lambda_1 - \lambda_3).\frac{r}{(n-r)}\int_{\theta}^{+\infty} P_r(x)dx + (\lambda_2 - \lambda_4)\int_{\theta}^{+\infty} P_{nr}(x)dx\right]$$

$$\int_{\theta}^{+\infty} P_r(x)dx = \beta_{j^*1}(\theta) + \beta_{j^*1}(\theta) + \alpha_{11} + \alpha_{12}\cdots\alpha_{cp1} + \alpha_{cp2}$$

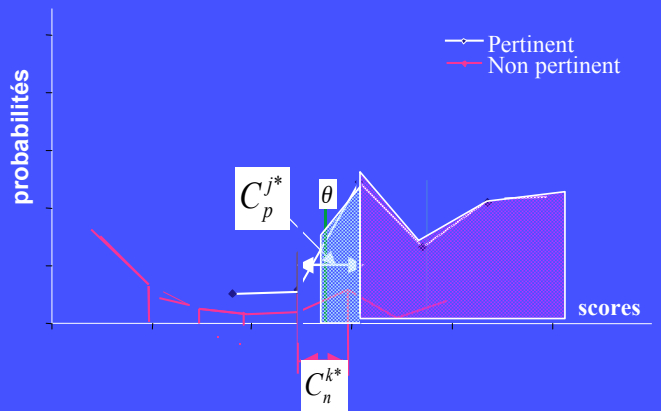
$$\int_{\theta}^{+\infty} P_{nr}(x)dx = \beta'_{j^*1}(\theta) + \beta'_{j^*1}(\theta) + \alpha'_{11} + \alpha'_{12}\cdots\alpha'_{cn1} + \alpha'_{cn2}$$



Adaptation de la fonction de décision (8/8)

3. Sélection du seuil optimal

$$\theta^* = \arg\max_{\theta} \left[\begin{array}{l} \frac{r.(\lambda_1 - \lambda_3)}{(n-r)} (\beta_{j^*1}(\theta) + \beta_{j^*2}(\theta)) + (\lambda_2 - \lambda_4)(\beta'_{k^*1}(\theta) + \beta'_{k^*2}(\theta)) \\ + \frac{r.(\lambda_1 - \lambda_3)}{(n-r)} \sum_{i>j^*} (\alpha_{i1} + \alpha_{i2}) + (\lambda_2 - \lambda_4) \sum_{i>k^*} (\alpha'_{i1} + \alpha'_{i2}) \end{array} \right]$$



Expérimentations

- **Collection de test (banc d'essai) proposée dans le cadre du programme TREC**
  - Un ensemble de documents issus du corpus Reuters (810.000)
  - 50 «topics» (numérotés de 101-150) utilisés comme profils
  - Une liste de documents pertinents par profil
  - À titre d'exemple :
    - Profil 126 contient 586
    - Profil 101 contient 308
    - Profil 137 contient 9

- **Mesure d'évaluation : Fonction d'utilité**

$$T11U = 2.R_+ - S_+$$

$$T11SU = \frac{2}{3} * \left(\max\left(\frac{T11U}{2R}, -0.5\right) + 0.5 \right)$$

Exemple de profil

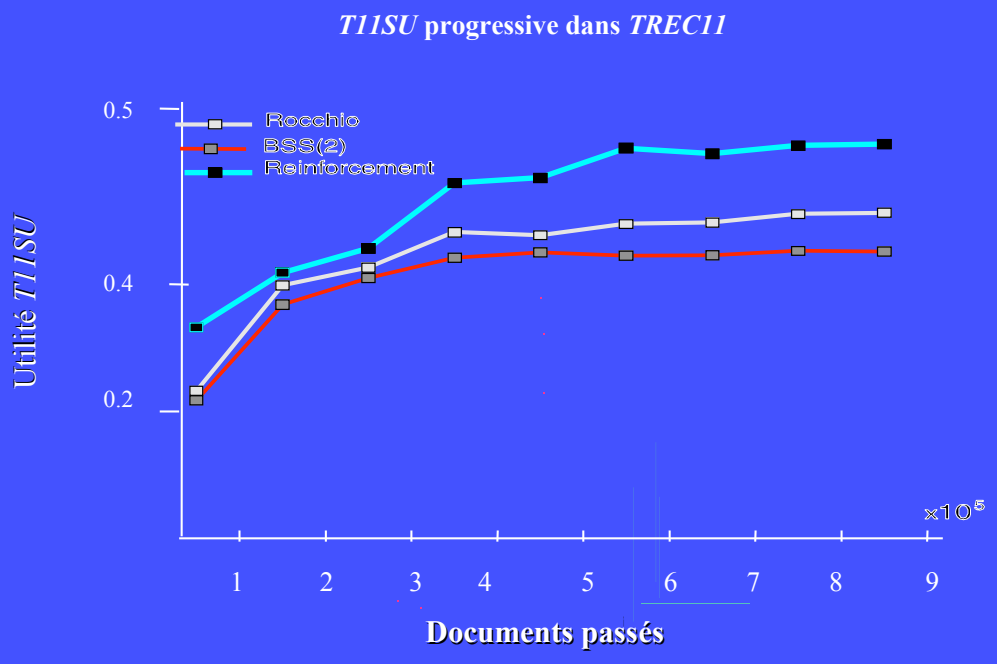
<num> Number: R101
<title> Economic espionage

<desc> Description:
What is being done to counter economic espionage internationally?

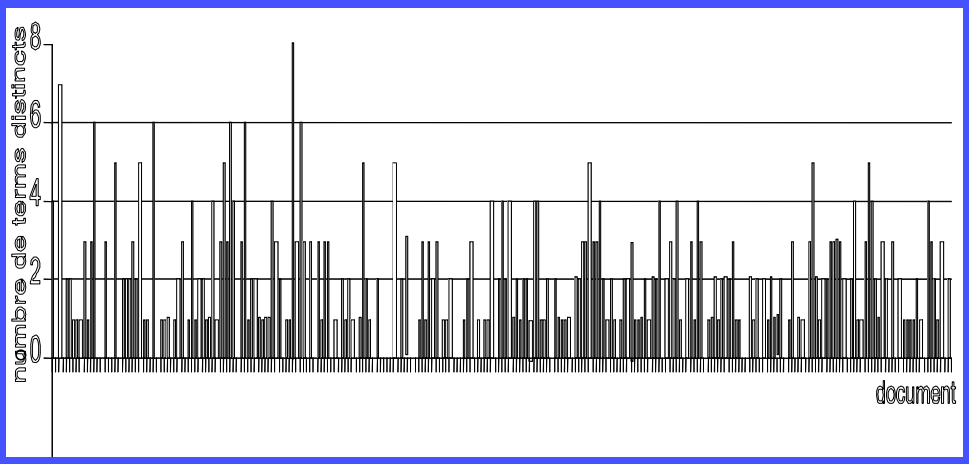
<narr> Narrative:
Documents which identify economic espionage cases and provide action(s) taken to reprimand offenders or terminate their behavior are relevant.
Economic espionage would encompass commercial, technical, industrial or corporate types of espionage. Documents about military or political espionage would be irrelevant.

ex. après apprentissage

Résultats : adaptation du profil



Différence entre les profils



Nombre de termes distincts entre deux profils appris par renforcement et Rocchio

Comparaison avec les résultats de TREC'2002

Participant	Moy. T11F	Moy. T11SU	Moy. T11U	Type apprent.
IRIT	0.462	0.474	69.04	
Chinese Academy of Sc.	0.427	0.475	60.76	Rocchio
Microsoft R. Cambridge	0.421	0.435	49.4	BM25
Tsinghua Univ.	0.417	0.395	43.78	Rocchio
Carnegie Mellon Univ.	0.41	0.447	51.3	Rocchio
KerMIT Consortium	0.376	0.459	59.04	
CLIPS IMAG Lab	0.369	0.424	44.12	R. Neurones
Fundan Univ.	0.346	0.397	23.8	Rocchio
Independent C.Lewis	0.318	0.293	14.08	
Queens College CUNY	0.196	0.154	-156.82	R. Neurones
Rutgers Univ.-Kantor	0.187	0.337	12.62	
Univ. of Iowa	0.174	0.333	6.64	
Johns Hopkins Univ.	0.104	0.342	5.5	
Univ. of Buffalo-Cedar	0.014	0.013	--	

Tab. Les participants à TREC-2002 : Filtrage adaptatif

Conclusion et perspectives

- **Dans le cadre de ce travail**
  - Profil utilisateur est vu de manière simpliste « liste de mots clés »
  - Seules deux processus du cycle de vie sont considérés
- **Objectif ACI-APMD**
  - Intégrer d'autres dimensions du profil
  - Intervenir sur tous les processus du cycle de vie de la requête.