Prise en compte de la structure des documents pour la découverte d'informations inattendues

François Jacquenet
Francois.Jacquenet@univ-st-etienne.fr

Christine Largeron
Christine.Largeron@univ-st-etienne.fr

Soutenu par le projet BINGO - ACI Masses de Données 2004-2007

Contexte

- Recherche d'informations nouvelles dans des textes [Bun et al.,01, Matsumura et al. 01, Liu et al. 01]
  - Program Topic Detection and Tracking, DARPA 96 [Allan et al.,98, Wayne, 98]
  - Challenge Novelty detection, TREC 2003 [Soboroff et al., 03]
  - Système UnexpectedMiner [Jacquenet et Largeron, PKDD 04, RNTI 06]
- Prise en compte de la structure du document [Dkaki et al., 04]

Plan

- Prise en compte de la structure du document
- Le système UnexpectedMiner
- Mesures d'inattendu
- Expérimentations
- Conclusion et perspectives

Prise en compte de la structure (1/3)

- Prétraitement : lemmatisation, mots vides
- Représentation vectorielle [Salton et Buckley, 88]
  - Un document j est représenté par un vecteur de poids

$$\overrightarrow{d_j} = \left(w_{1j}, w_{2j}, \ldots, w_{ij}, \ldots, w_{pj}\right)$$

  - Le poids w_{ij} du terme t_i dans le document d_j est évalué par la formule TF.IDF

$$w_{ij} = \begin{cases} 0 & \text{si } f_{ij} = 0 \\ tf_{ij} \times idf_i & \text{sinon} \end{cases}$$

$$tf_{ij} = \frac{f_{ij}}{\max_l f_{lj}}$$

$$idf_i = \log \frac{N}{n_i}$$

Prise en compte de la structure (2/3)

- Document composé de k parties
- Coefficients de structuration CS_l, l = 1,…,k

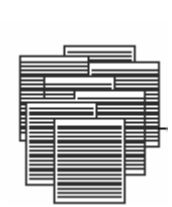
$$\sum_{l=1,k} cs_l = 1$$

- Le poids w_{ij} du terme t_i dans d_j est défini par

$$tf_{ij}^{l} = \frac{f_{ij}^{l}}{\max_h f_{hj}} \qquad w_{ij} = k \sum_{l=1}^{k} cs_l . tf_{ij}^{l} . idf_i$$

Prise en compte de la structure (3/3)

- Comment choisir les coefficients ?
- Apprentissage à partir d'un échantillon de textes



UnexpectedMiner
$C_i = (C_{i1}, \ldots, C_{ik})$

Rang	Document
1	d_m
…	…
r	d_j
…	…
|E|	d_n

Minimisation de la fonction d'erreur $F(C_l)$

Document fourni à tort

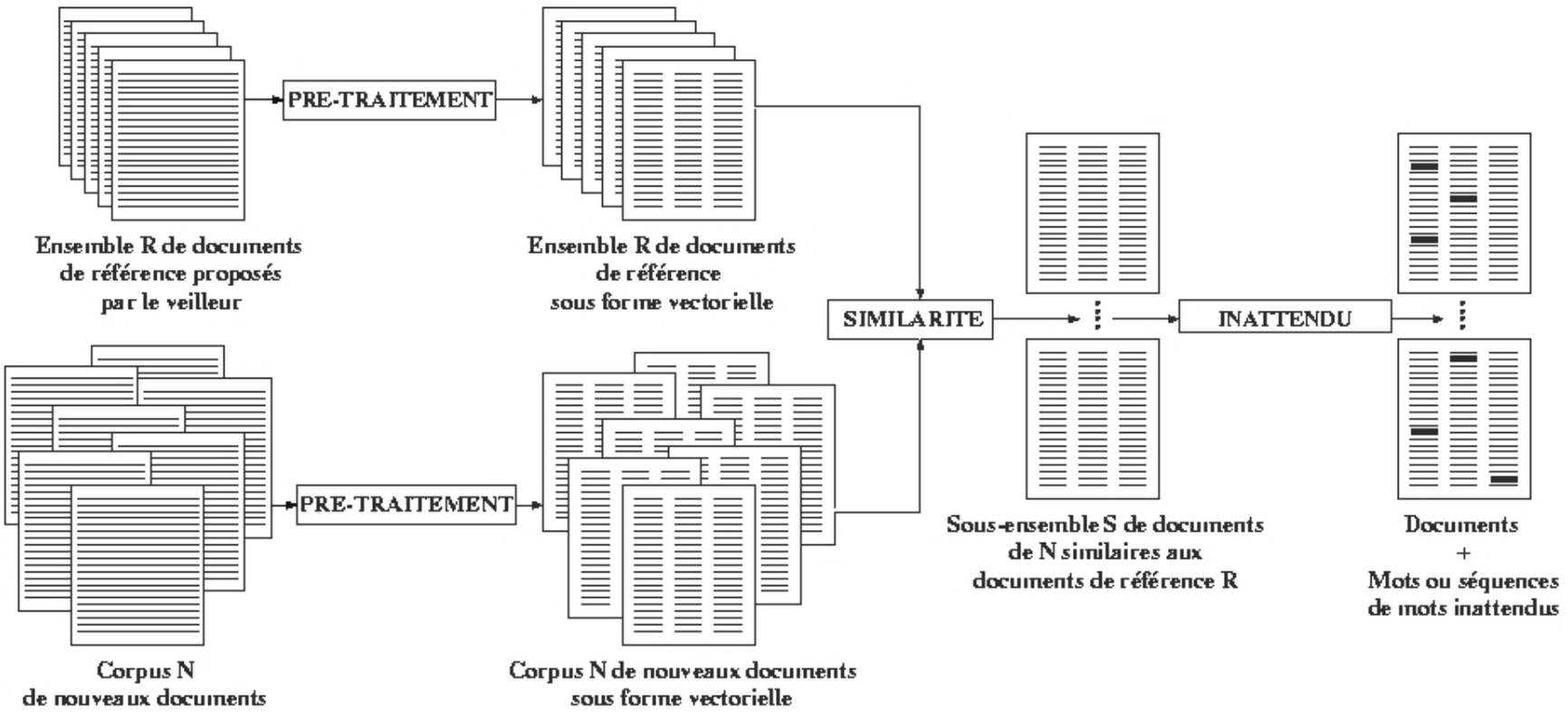
$$f_{C_l} = \sum_{j=1}^{|E|} \frac{1}{r}[j] \text{ où } [j] = \begin{cases} 1 \ si \ d_j \ est \ erroné \\ 0 \ sinon \end{cases}$$

Rang des erreurs	Fonction d'erreur
1, 2, 3	1,83
3, 4, 7	0,73
8, 9, 10	0,34

- Optimisation de la fonction erreur par recuit simulé

Le système UnexpectedMiner

Architecture



Mesures d'inattendu

Mesure de WebCompare [Liu et al. 2001]

- $$M1(d_j) = \frac{\sum_{i=1}^{m} U^1_{ijc}}{m}$$

- $$U^1_{i,j,c} = \begin{cases} 1 - \frac{tf_{i,c}}{tf_{i,j}} \text{ si } tf_{i,c} / tf_{i,j} \leq 1 \\ 0 \text{ sinon} \end{cases}$$

Prise en compte du pouvoir discriminant idf

[Jacquenet et al. , 2004]

- Somme des poids w_{ij} des termes t_i dans le document d_j

$$M2(d_j) = \sum_{i=1}^{m} w_{ij} \qquad w_{ij} = tf_{ij} \times idf_i$$

- Poids maximum observé dans d_j

$$M3(d_j) = \operatorname{Max}_i w_{ij}$$

Expérimentations

Protocole

- Corpus
  - Documents de référence (R) : **18**
  - Nouveaux documents (N) : **223**
  - **40** Documents similaires (S) à ceux de R dont **18** inattendus
- Evaluation du module d'inattendu
- Evaluation du système
- Précision et rappel, fonction d'erreur f

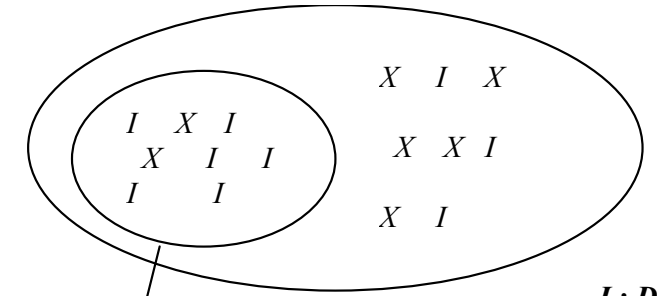
Précision et rappel [Swets, 63]

- Précision

  Pourcentage de documents extraits par le système qui sont inattendus

- Rappel

  Pourcentage de documents inattendus trouvés par le système dans le corpus de nouveaux documents N



Précision : 6/8

Rappel : 6/9

I : Documents Inattendus
X : Autres documents

Fonction objectif

- Test du module inattendu

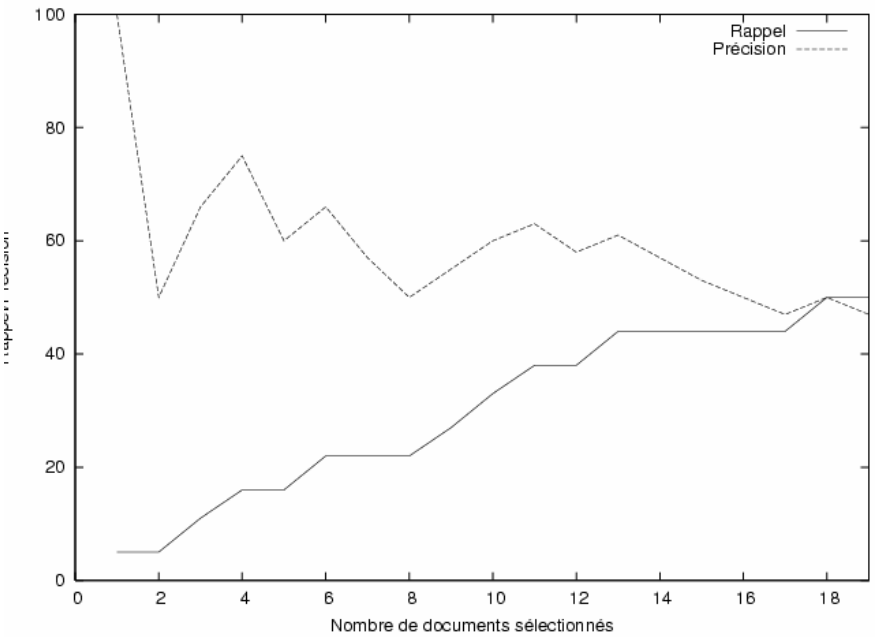
Mesure	Fonction d'erreur
WebCompare	2,07
Somme des poids	0,92
Poids maximum	0,59

- Test du Système UnexpectedMiner complet

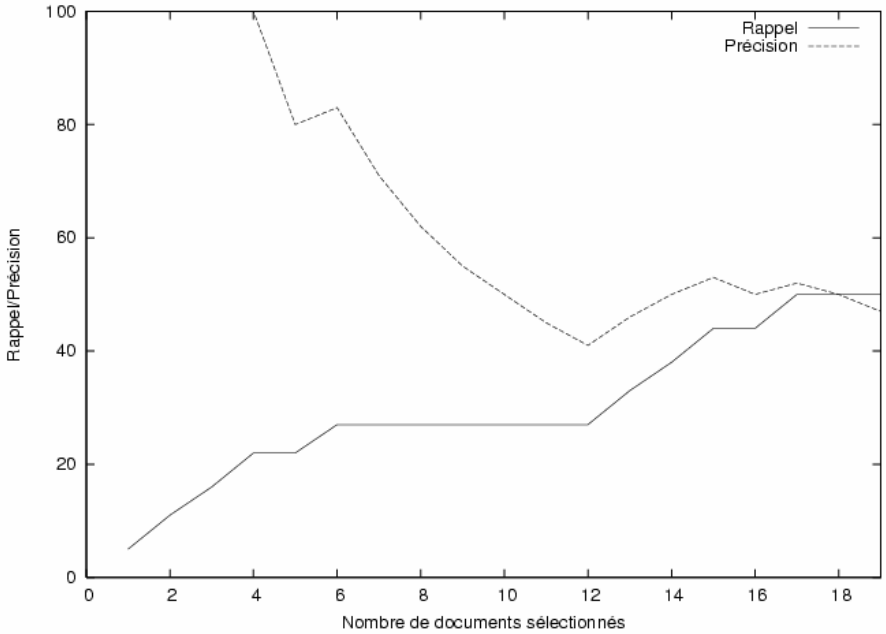
Mesure	Fonction d'erreur
WebCompare	2,71
Somme des poids	1,9
Poids maximum	1,76

M2 : Mesure de la somme des poids

Sans structure

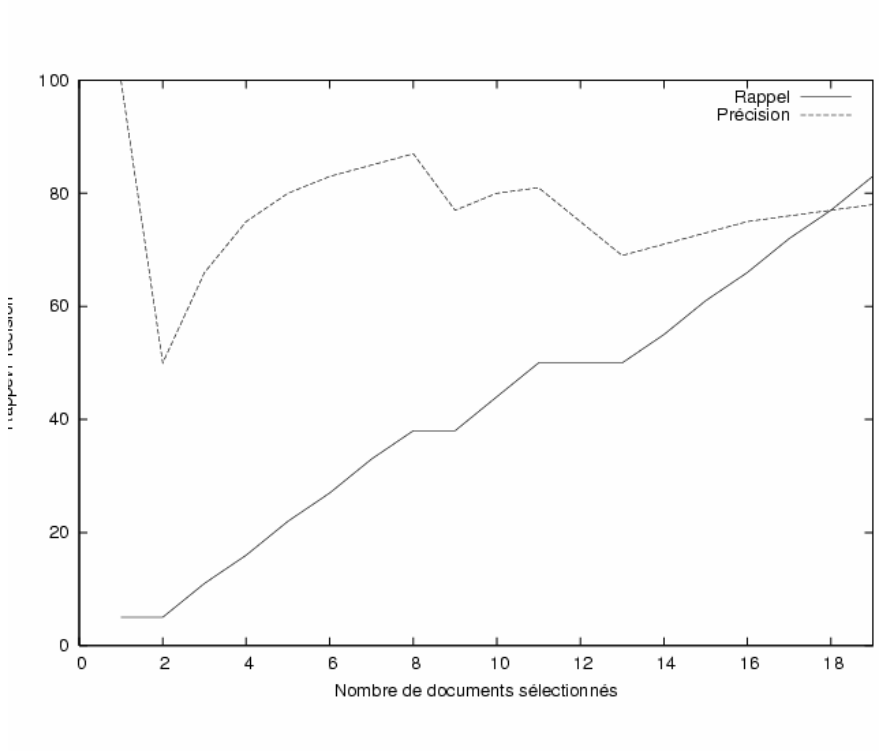


Avec structure

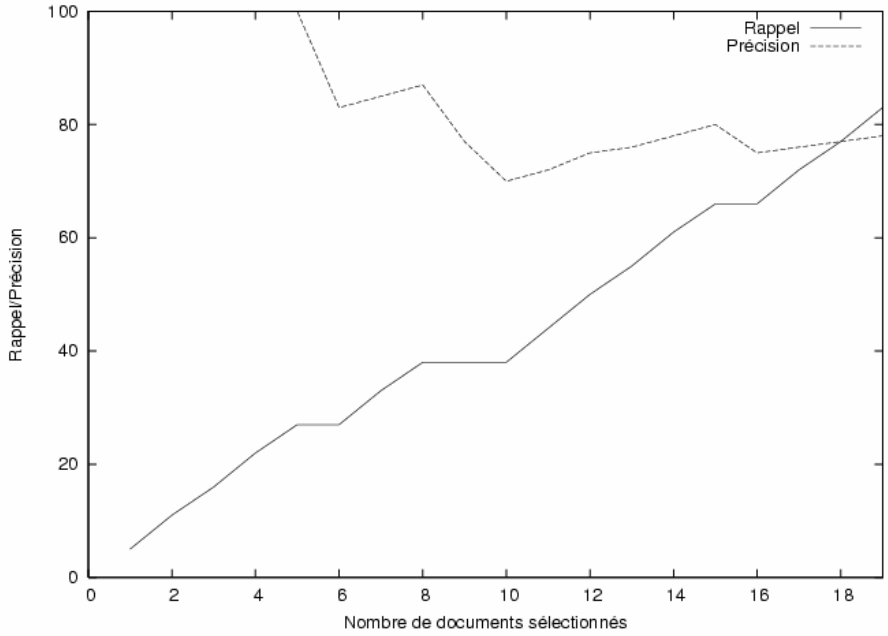


M3 : Mesure du poids maximum

Sans structure



Avec structure



Conclusion et perspectives

- Intérêt de la structure des documents pour la recherche de nouveauté

- Calcul automatique des poids accordés aux parties par apprentissage

- Représentation du document

- Similarité entre documents

Merci

François Jacquenet
Francois.Jacquenet@univ-st-etienne.fr

Christine Largeron
Christine.Largeron@univ-st-etienne.fr

Similarité

- Extraire du corpus des N nouveaux documents, les plus proches des documents de référence R
- Distance du cosinus

$$s_{jk} = \frac{\vec{d}_j.\vec{d}_k}{|\vec{j}| \times |\vec{k}|} \; où \; \vec{d}_j.\vec{d}_k = \sum_i w_{ij} \times w_{ik}$$

- Similarité moyenne s_j d'un nouveau document d_j de N avec les documents de R

$$s_j = \frac{1}{|R|} \sum_{k=1}^{|R|} s_{jk}$$

- Tri par ordre de similarité décroissante : $S \subset N$

Mesure de WebCompare [Liu et al. 2001]

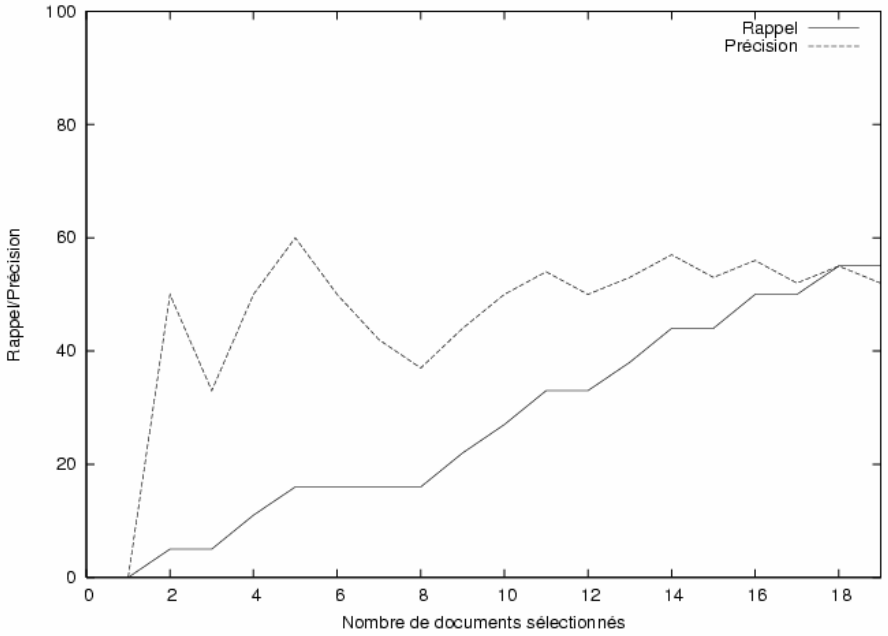
$$M1(d_j) = \frac{\sum_{i=1}^{m} U^1_{ijc}}{m}$$

$$U^1_{i,j,c} = \begin{cases} 1 - \dfrac{tf_{i,c}}{tf_{i,j}} \text{ si } tf_{i,c} / tf_{i,j} \leq 1 \\ 0 \text{ sinon} \end{cases}$$

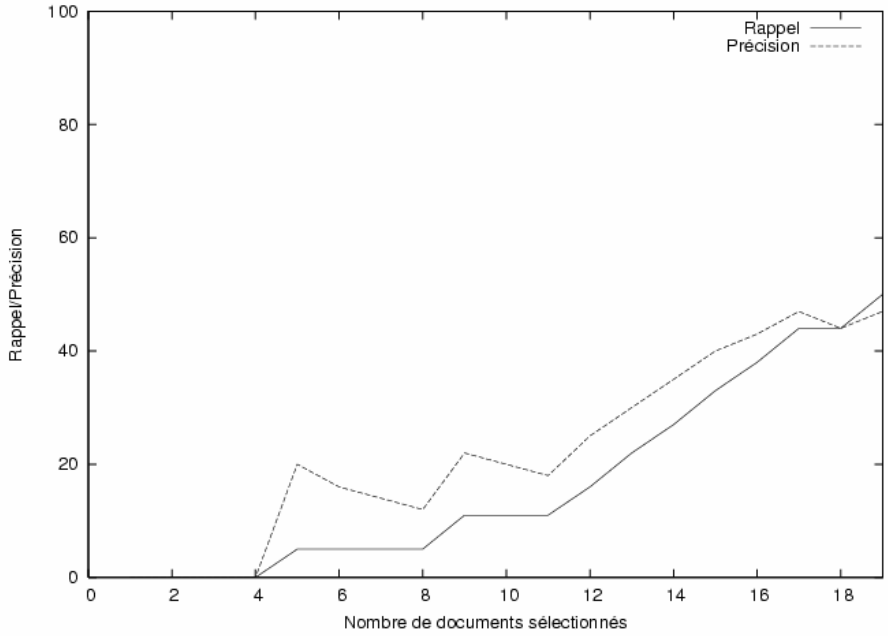
Document d_j		Set $R \cup S - \{d_j\}$		
Term t_i	Frequency	Term t_i	Frequency	$U_{i,j,c}$
...	...	...	...	...
mibing	1	mibing	0	1
...	...	...	...	...
boosting	4	boosting	0	1
...	...	...	...	...

Mesure de WebCompare

Module inattendu

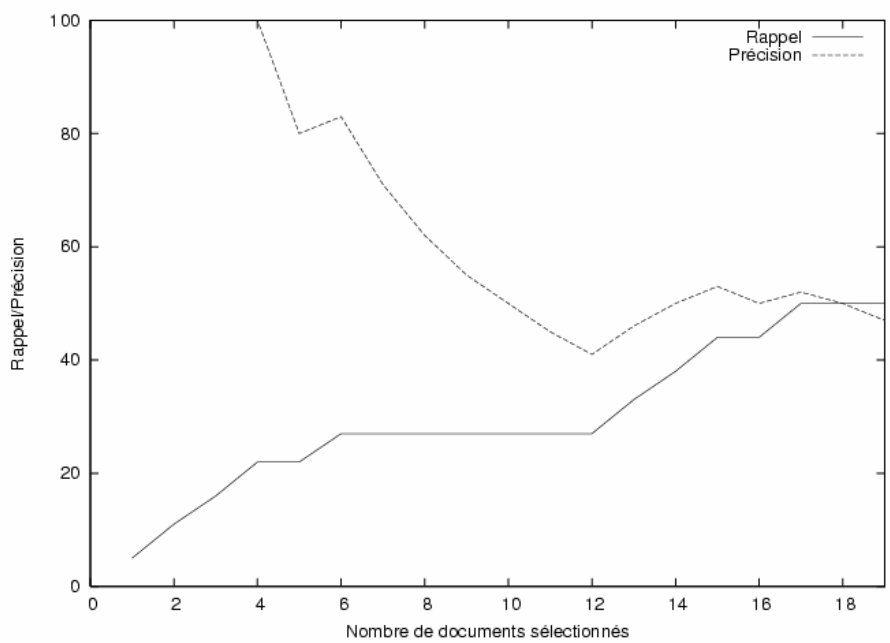


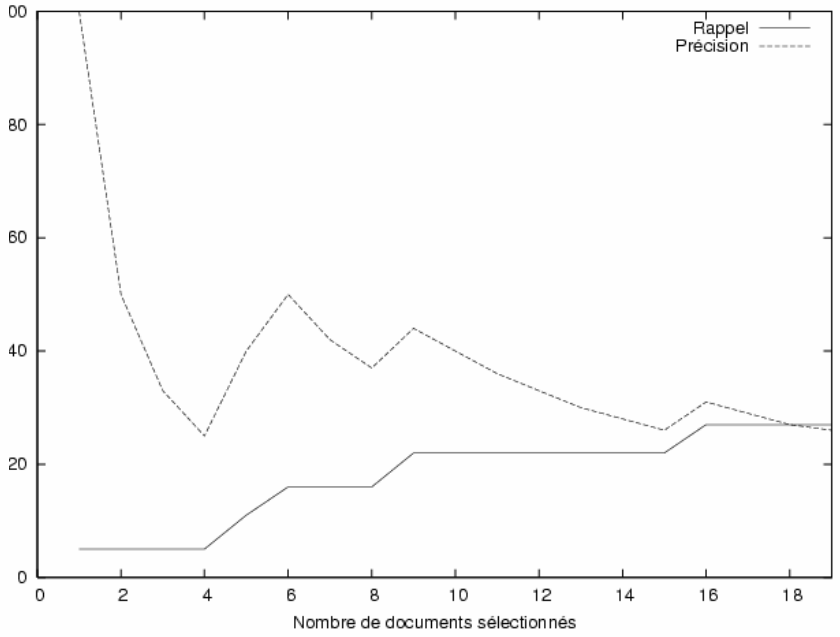
Système UnexpectedMiner



M2 : Mesure de la somme des poids

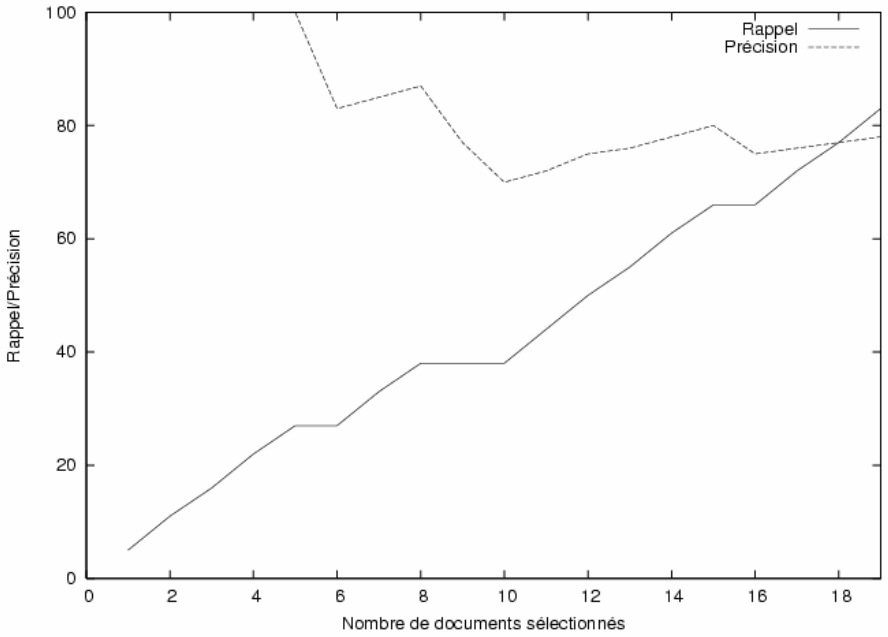
Module inattendu Système UnexpectedMiner





Mesure du poids maximum

Module inattendu



Système UnexpectedMiner