Annexe : Évaluation d'un système de recherche d'information

Indexation de données multimédias

A. ELHASSOUNY

GL
ENSIAS

BDMM

Mesures et évaluation

- Le but de système d'indexation est de trouver les documents pertinents à une requête, et donc utiles pour l'utilisateur

- … et pour mesurer la performance d'un système on peut utiliser soit les mesures subjectives ou objectives

  - Mesures objectives, subjectives

  - Évaluation d'un système de recherche d'information

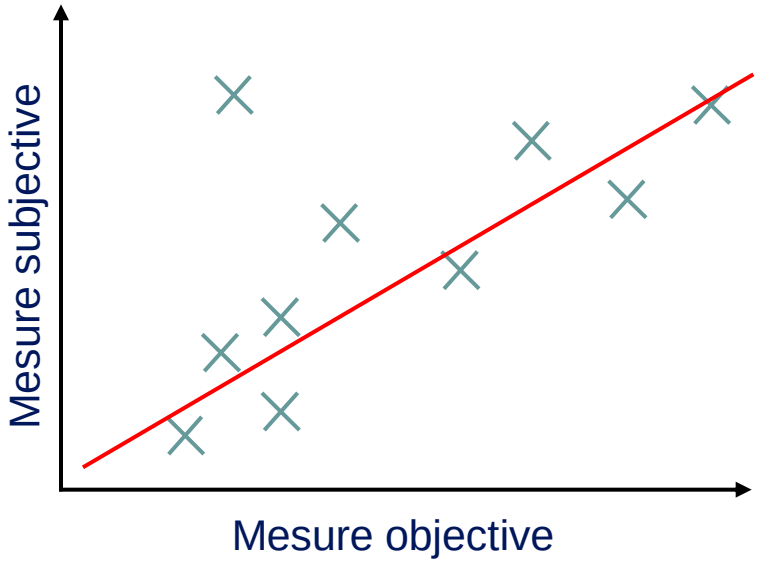
Mesures subjectives

- Pour beaucoup d'applications, le but est de maximiser l'espérance de satisfaction de l'utilisateur.
  - Seule une mesure subjective par l'utilisateur lui-même permet d'optimiser pleinement ce critère

- Difficultés :
  - L'avis d'un utilisateur peut varié et n'instancie pas un ordre total
  - Deux utilisateurs distincts ne portent pas le même jugement
  - !!! Le coût !!!

Mesures objectives/subjectives

- Idée : Chercher une mesure objective qui modélisera la mesure subjective
  - Pour une application particulière (ou dans un domaine)



Évaluation d'un système de recherche d'information

- Évaluation directe avec les utilisateurs
  - Complexe à mettre en place
  - Coûteuse

- Évaluation à partir d'une vérité-terrain : Mesures de performances
  - Courbes de précision/rappel
  - Matrices de confusion
  - Courbes ROC (Receiver Operating Characteristics)
  - Rapidité de convergence de la recherche itérative

Pré-requis pour l'évaluation

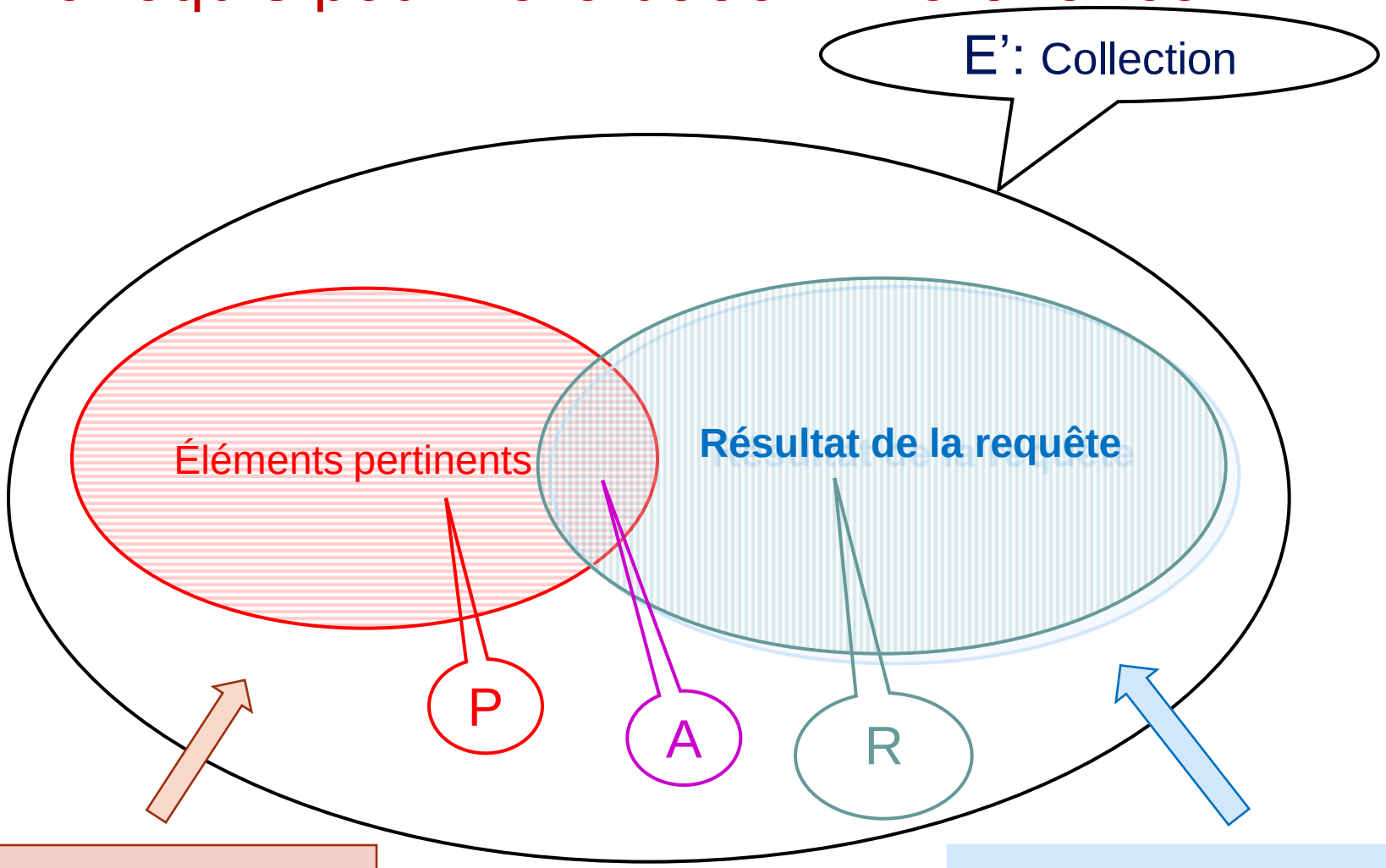
- Avoir à disposition
  - un ensemble de test (base dans laquelle on recherche)
  - un ensemble de requêtes non incluses dans la base de test
  - une vérité terrain (ground truth) pour chaque couple (requête, éléments de la base) qui répond à la question : est-ce que l'élément de la base est pertinent pour la requête considérée ?
- Remarques
  - pour comparer deux méthodes, mêmes ensembles de test et requêtes doivent être utilisés
    - bases de tests partagées par les chercheurs du domaine
    - compétition avec introduction de nouvelles bases de test
  - la taille de ces ensembles doit être suffisamment grande pour diminuer la variance de l'évaluation

Pré-requis pour l'évaluation : Pertinance

- La pertinence comme mesure pour la recherche : chaque document est classifié pertinent ou non pertinent pour une requête

  - Cette classification est effectuée manuellement par des **«experts»**
  - La réaction du système à la demande de recherche sera par rapport à cette classification
  - Comparer **la réponse obtenue** avec **le résultat "idéal"**

Pré-requis pour l'évaluation : Pertinance



Pré-requis pour l'évaluation : Pertinance

- Soit E un ensemble d'objets (l'ensemble des textes, images, vidéos) muni d'une distance **r** telle que
  - Soient **x, y ∈ E, r(x,y) = 0** si **y** est pertinent pour **x**
    r(x,y) = 1 sinon
- Soit **q(,)**, distance instancie la vérité terrain
  - Exemple : **x et y** sont 2 images
    q(x,y) = 0 si **x et y** se ressemblent,
    q(x,y) = 1 sinon
- Soit un ensemble **E' ⊂ E, et x : x ∈ E et x ∉ E'**
  - **E' : ensemble dans lequel on effectue la recherche**
  - **x : la requête**

Précision/rappel

- Le système de recherche est **paramétré** pour retourner plus ou moins de résultats
  - la quantité d'objets retournés varie entre 1 et #E'
  - plus on retourne de résultats, plus on a de chance de retourner toutes les objets pertinents de la base
  - en général, moins on en retourne, plus le taux d'objets retournés et qui sont pertinents est élevé
- Ces deux notions sont couvertes par les mesures de **précision** et de **rappel**

Précision/rappel

- Soit R l'ensemble des résultats retournés, de cardinal #R
- Soit P l'ensemble des éléments pertinents dans E' pour x, c-a-d

$$P = \{ y \in E' / q(x,y) = 0 \}$$

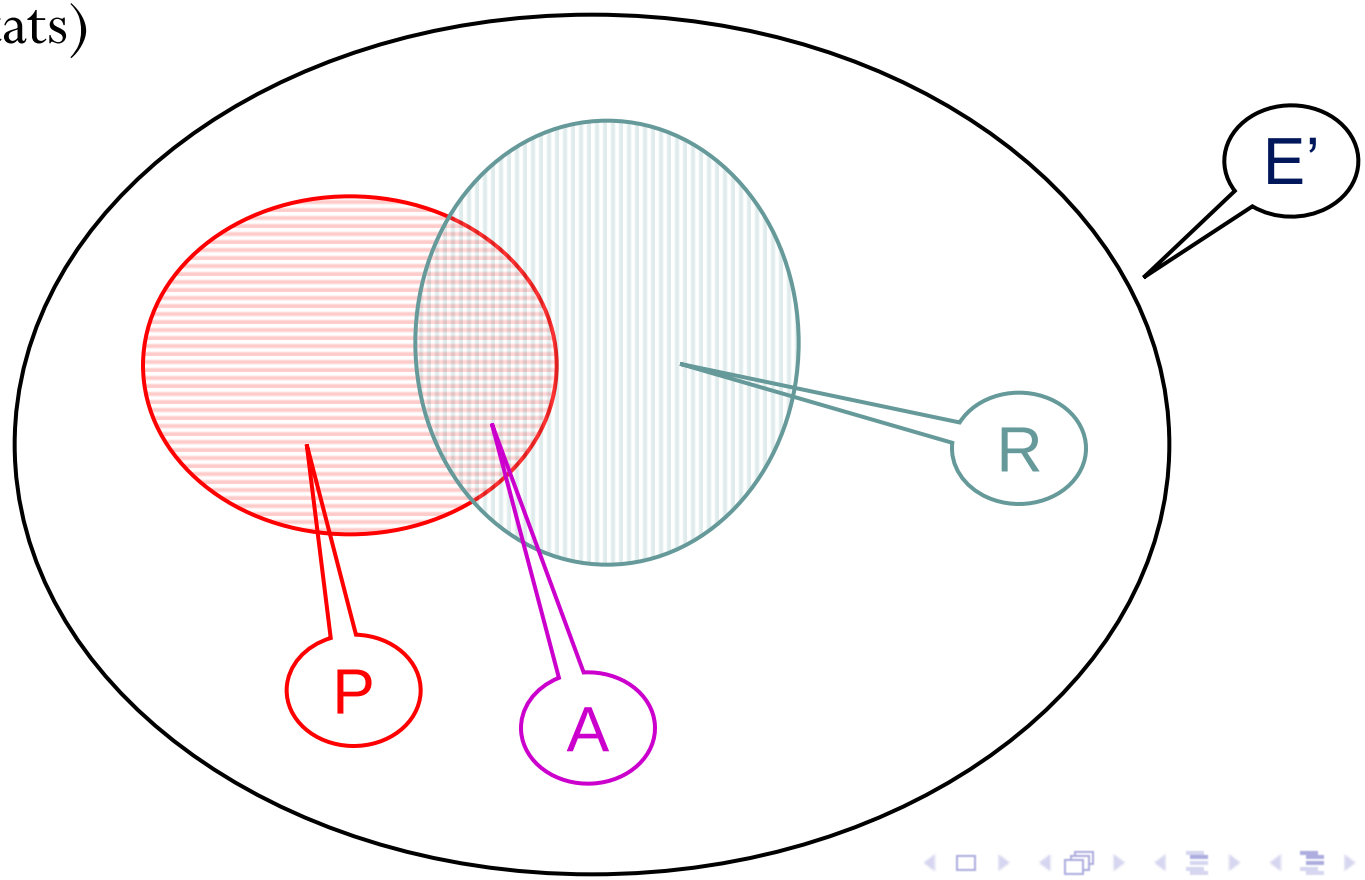
- Soit A l'ensemble des résultats retournés et qui sont pertinents

$$A = \{ y \in R / r(x,y) = 0 \text{ et } q(x,y)=0 \}$$

- **la précision = #A / #R =** est le taux d'éléments qui sont pertinents parmi ceux qui sont retournés par le système
- **le rappel = #A / #P =** est le taux d'éléments qui sont pertinents qui sont retournés par le système
- La performance du système peut être décrite par **une courbe précision/rappel**

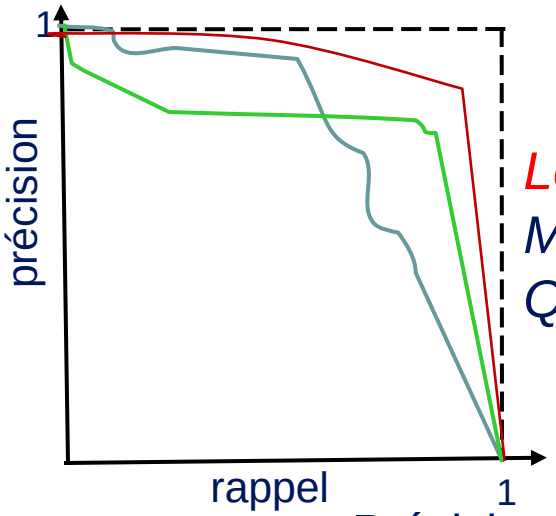
Précision/rappel

- Remarques :
  - P est indépendant de la requête.
  - R varie en fonction de la paramétrisation (qui retourne + ou – de résultats)



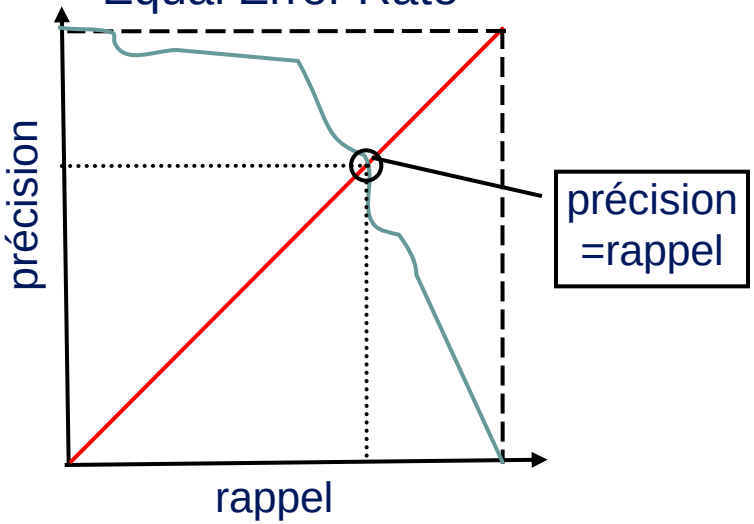
Courbe précision/rappel

- La courbe précision/rappel n'instancie par un ordre total

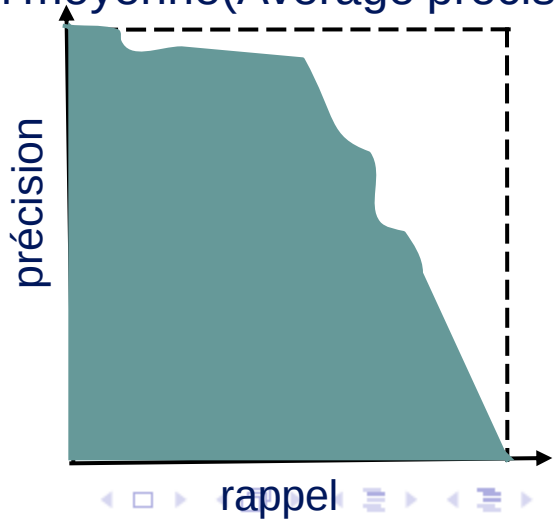


Le systeme rouge est le meilleur
Mais pour le vert et le bleu :
Quel est le mieux?

Equal Error Rate



Précision moyenne(Average precision)



Courbes ROC (Receiver operating characteristic)

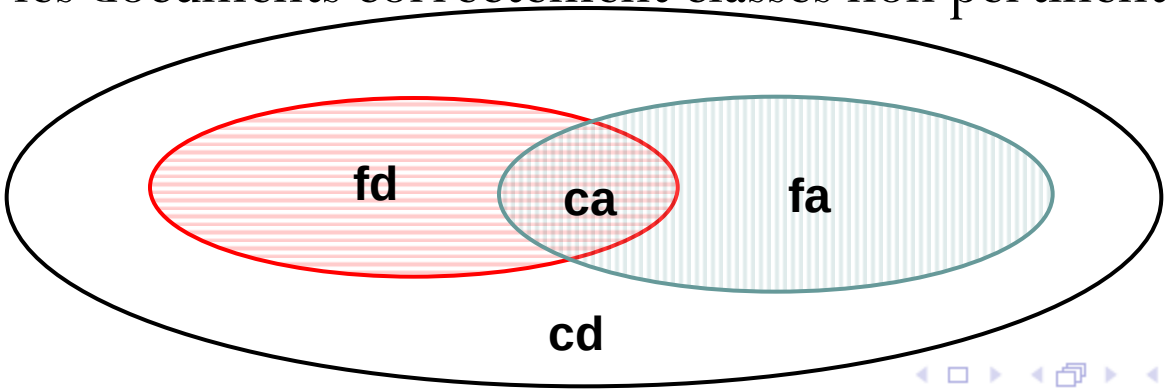
- Courbes ROC : sensitivité, spécificité
  - Soit une vérité terrain q(.,.)
  - Réponse du système à une requête x, r(x,y)=0 si y est retourné (objet considéré pertinent), r(x,y)=1 sinon

Système \ Vérité terrain	Pertinent (p)	non pertinent (n)
Pertinent (p')	**Vrai positif (vp)** q(x,y)=0 r(x,y)=0	**Faux positif (fp)** q(x,y)=1 r(x,y)=0
Non pertinent (n')	**Faux négatif (fn)** q(x,y)=0 r(x,y)=1	**Vrai négatif (vn)** q(x,y)=1 r(x,y)=1

  - **Sensitivité =rappel= vp / (vp + fn)**
  - **Spécificité: taux de faux positifs = fp / (fp + vn) = 1 - spécificité**
- Courbe ROC : rappel en fonction du taux de faux positifs

Courbes ROC (Receiver operating characteristic)

- Faux positifs (False Positives): Fausses alarmes **(False alarmas) : fa**
  - Documents non pertinents, classés comme pertinente par le système
- Faux Négatifs **(False Negatives) : False dismissals : fd**
  - Documents pertinents classés par le système comme non pertinents
- Vrai positifs **(correct alarms) : ca**
  - Tous les documents correctement classés pertinente par le Système
- Vrai négatifs **(correct dismissals): cd**
  - Tous les documents correctement classés non pertinents par le système



Courbes ROC : Area under Curve (AUC)

- AUC : Mesure de performance calculée à partir de la courbe ROC
  - Exemple pour mesure la pertinence d'un test médical (voir http://gimm.unmc.edu/dxtests/roc3.html)

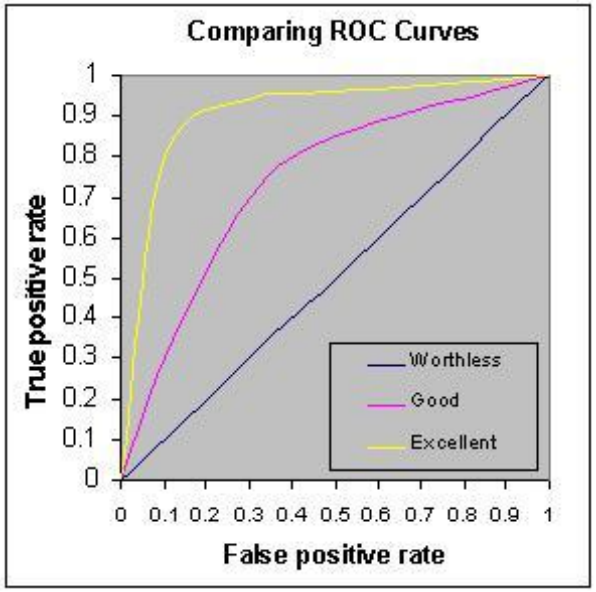
0.90-1.00 Excellent

0.80-0.90 Bon

0.80-0.70 Passable

0.60-0.70 Pauvre

0.50-0.60 Mauvais



- Courbe ROC (Receiver Operating Characteristics) : sensibilité vs. 1-spécificité pour différents seuils de décision

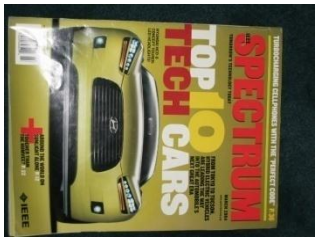
Et la pertinence ?

- La pertinence d'un système (pour une paramétrisation donnée) est le taux d'objets qui sont correctement jugée, c-à-d
  - **pertinence = (vrais positifs + vrais négatifs) / taille de la base**
- Intérêt d'avoir des courbes (précision/rappel et ROC) pour l'évaluation
  - dépend de l'utilisation : certains utilisateurs **cherchent la précision** (ex: requête sur Google), d'autres **un grand rappel possible** (recherche de contenu piraté)
- En effet, il est facile d'avoir 100% de rappel, il suffirait de donner toute la base comme la réponse à chaque requête. Cependant, la précision dans ce cas-ci serait très basse. De même, on peut augmenter la précision en donnant très peu de documents en réponse, mais le rappel souffrira. Il faut donc utiliser les deux métriques ensemble

Exercice : système de recherche d'objets

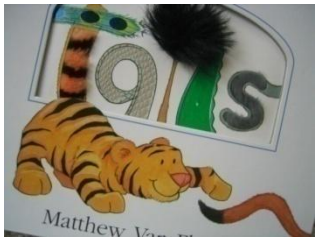
- Pour la requête et les résultats triés suivants : tracer les courbes précision/rappel et ROC.

1 2 3 4 5

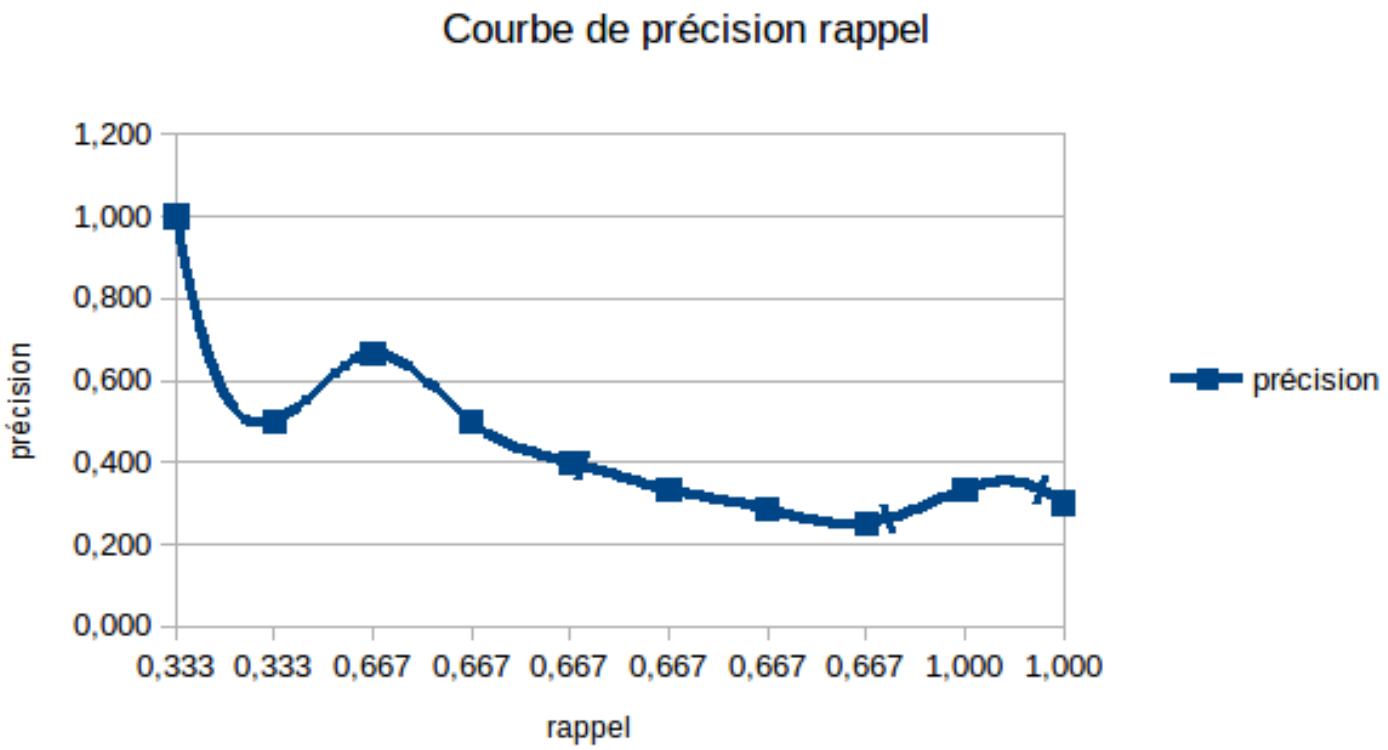





6 7 8 9 10


Exercice : système de recherche d'objets



Mesures et protocole d'évaluation : conclusion

- Difficulté de trouver une bonne mesure
  - elle doit être adaptée à ce que l'on compare
  - elle doit répondre à l'objectif recherché

- Évaluation d'un système de recherche multimédia
  - méthodes identiques à celles utilisées en texte
  - utilisation de courbes plutôt que de scalaires (peuvent être interprétées en fonction du besoin).

Conclusion

- Indexation des images et vidéos : problème non résolu
- Experts issus de domaines variés (informatique, traitement de l'image, psycho visuel, apprentissage automatique, ...)
- Deux axes à étudier simultanément :
  - Techniques d'analyse d'image donc d'extraction et de comparaison de l'information
  - Pertinence de l'information pour un utilisateur
- Produits commerciaux encore basiques ...

Pour plus d'info.

- Introduction to Information Retrieval, C. D. Manning, P. Raghavan and H. Schütze, Cambridge University Press, 2008
  - Chapitre 8
- http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html