

FRI : Fondements en Recherche d'Information

Cours 4 : Evaluation

Céline Hudelot, Maître de conférences, Centrale Paris

2015-2016

Avant propos

Lectures conseillées

- Chapitre 8 du livre *Introduction to Information Retrieval*.
<http://nlp.stanford.edu/IR-book/>
- Chapitre 8 du livre *Search Engines. Information Retrieval in Practice*.
<http://ciir.cs.umass.edu/irbook/>

Plan

- 1 Introduction
 - Evaluation par banc d'essais
 - Evaluation par étude d'utilisateurs
 - Evaluation en ligne
- 2 Requête vs Besoin d'Information
- 3 Evaluation de résultats de recherche sans classement
- 4 Evaluation de résultats de recherche avec classement
- 5 Les collections pour l'évaluation
- 6 Présentation des résultats
- 7 Conclusion

Introduction

Pourquoi évaluer ?

- Beaucoup de modèles de recherche différents, de nombreux algorithmes. Quel est le meilleur pour mon application ?
- Quel est le meilleur composant pour :
 - ▶ La fonction de classement : cosinus, produit scalaire, ...
 - ▶ La sélection des termes d'indexation (suppression des stop-word, tokenization, ...)
 - ▶ Mesure pour l'établissement des poids (tf, tf-idf, ...)
- Une *science expérimentale* : en RI, on ne peut pas prouver que A est meilleur que B sans expérimentations.

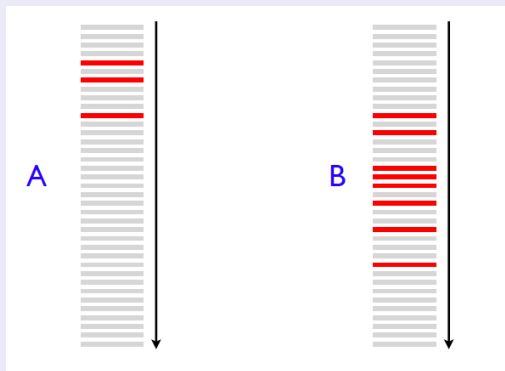
Efficacité de la recherche

- En réponse à une requête, un système de RI cherche dans une collection de documents et retourne une liste ordonnée de réponses
 - ▶ Résultat : liste avec un rang
 - ▶ Stratégie et algorithme de recherche
- Objectif : mesurer la qualité de la liste ordonnée obtenue
 - ▶ Une meilleure stratégie amène à une meilleure liste
 - ▶ Une meilleure liste répond plus aux besoins de l'utilisateur

Efficacité de la recherche

Meilleur classement ?

Question difficile même en connaissant les documents pertinents.



Introduction

Evaluation en RI

- Comment se comporte le système ?
- Plusieurs niveaux :
 - ① Traitement : complexité en temps et en espace
 - ★ Rapidité de l'indexation : débit (*e.g. nombre de bytes par heure*).
 - ★ Rapidité de la recherche : latence (*e.g. temps de la réponse en secondes*)
 - ★ Coût par requête (*en dollars*).
 - ② Recherche : efficacité de la recherche
 - ③ Système : satisfaction de l'utilisateur
 - ★ Comment la quantifier ?

Un bon moteur de recherche ?

- Un moteur **rapide** : analyse rapide de la requête, recherche rapide dans l'index et tri rapide des résultats.
- Un moteur **complet** et **à jour** : traitement de tous les documents et ajouts de nouveaux documents.
- Un langage de requêtes **simple** et **expressif**.
- Une interface sympa ?
- Sa gratuité ?
- mais surtout sa **pertinence**.

Un bon moteur de recherche ?

Pertinence

- Les résultats doivent satisfaire le besoin d'information de l'utilisateur (difficile à mesurer mais pas indépendant des autres points)
- Une notion qui dépend de l'utilisateur et de ses besoins.

Besoins de l'utilisateur

Des besoins divers selon le type d'utilisateur et d'utilisation

• Types d'utilisateur

- ▶ Utilisateur d'un moteur de recherche WEB (*succès : l'utilisateur a trouvé ce qu'il cherchait*)
- ▶ Entreprise qui veut faciliter l'accès à ses données (publicité) (*succès : nombre de clics sur la publicité*)
- ▶ Vendeur sur un site de e-commerce (*succès : achat par l'utilisateur*)
- ▶ ...

• Types d'utilisation

- ▶ Recherche pour une compréhension : ex : *Effet du réchauffement de la planète sur les glaciers ?*
- ▶ Recherche d'un fait : ex : *Capitale du Brésil ?*
- ▶ Recherche d'information liée à une autre information : ex : *Etat de l'art sur les RI ?*
- ▶ ...

Notion de pertinence

Notion de pertinence

La satisfaction de l'utilisateur va être *mesurée* par la pertinence des résultats de la recherche

Comment mesure t-on la pertinence ?

- Notion difficile à définir
- Un document pertinent est un document utile dans le contexte d'un besoin (une requête)
- Problème des vraies collections de documents : impossible de connaître totalement l'ensemble des documents pertinents
- Les modèles de recherche ont une notion de pertinence :
 - ▶ Modèle booléen : satisfaisabilité d'une expression logique du premier ordre
 - ▶ Modèle probabiliste : $P(\text{relevance} \mid \text{requête}, \text{document})$
 - ▶ Modèle vectoriel : similarité

Notion de pertinence

Comment mesurer la pertinence ?

- Moteur de recherche
 - ▶ L'utilisateur clique sur certains liens et pas d'autres.
 - ▶ L'utilisateur retourne sur le moteur.
 - ▶ ...
- Site de e-commerce
 - ▶ Achat par l'utilisateur.
 - ▶ Achat rapide
 - ▶ Proportion de visiteurs qui achètent
 - ▶ ...
- SI d'entreprise
 - ▶ Productivité de l'utilisateur
 - ▶ ...

Mesure de la pertinence pour l'évaluation

3 grandes approches

- Evaluation par banc d'essais.
- Evaluation par des études sur les utilisateurs
- Evaluation en ligne.

Plan

- 1 Introduction
 - Evaluation par banc d'essais
 - Evaluation par étude d'utilisateurs
 - Evaluation en ligne
- 2 Requête vs Besoin d'Information
- 3 Evaluation de résultats de recherche sans classement
- 4 Evaluation de résultats de recherche avec classement
- 5 Les collections pour l'évaluation
- 6 Présentation des résultats
- 7 Conclusion

Evaluation par banc d'essais

Méthodologie classique

Elle nécessite trois éléments :

- Un banc d'essais (benchmark) sur la collection de documents
- Un banc d'essais de requêtes.
 - ▶ Souvent une description plus complète du besoin d'information est associé à la requête.
- Une mesure (souvent binaire) pour chaque paire (document, requête) faite par des humains
- Les systèmes sont évalués sur leur capacité à retrouver les documents pertinents pour ce banc de requêtes.
 - ▶ **Métrique d'évaluation** : une mesure qui quantifie la qualité d'un classement.

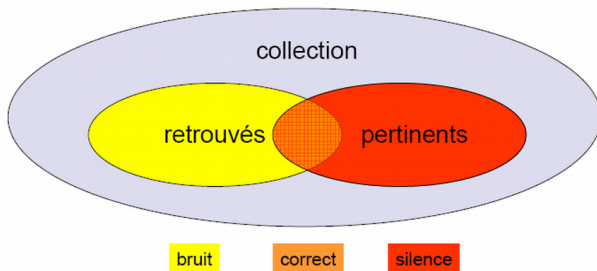
Une méthodologie souvent critiquée pour son manque de réalisme mais très utilisée en RI

Evaluation par bancs d'essai : métrique d'évaluation

Rappel

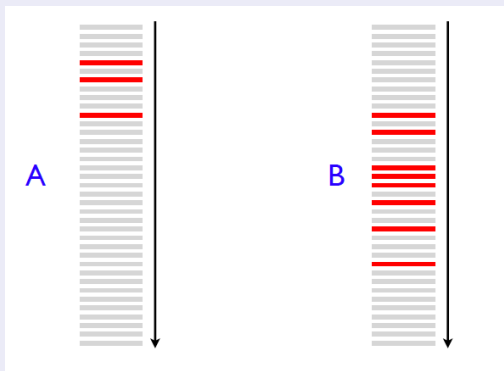
Rappel et précision

- Rappel : nombre de documents pertinents retrouvés par rapport au nombre de documents pertinents
- Précision : nombre des documents pertinents retrouvés par rapport au nombre de documents retrouvés



Evaluation par bancs d'essai : métrique d'évaluation

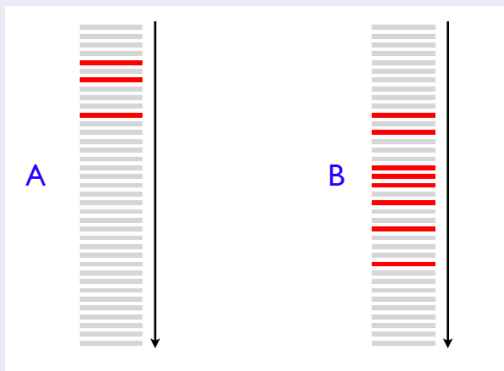
Quel est le meilleur classement ?



- Rang du premier document pertinent.

Evaluation par bancs d'essai : métrique d'évaluation

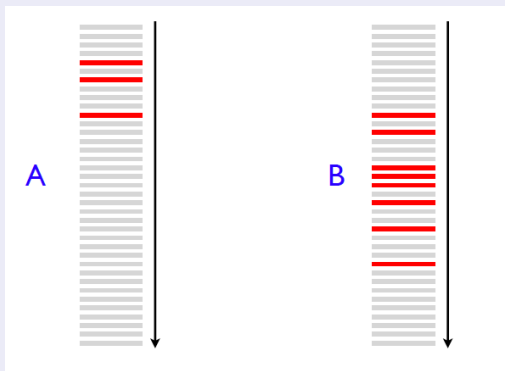
Quel est le meilleur classement ?



- Précision au rang 10

Evaluation par bancs d'essai : métrique d'évaluation

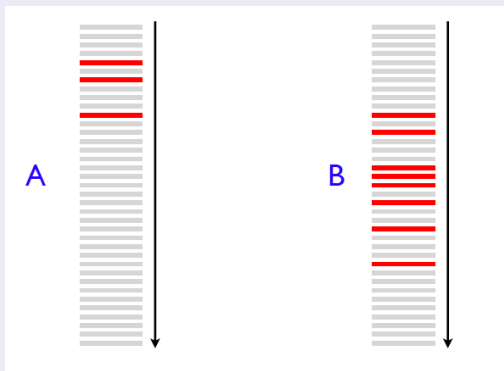
Quel est le meilleur classement ?



- Précision au rang 1

Evaluation par bancs d'essai : métrique d'évaluation

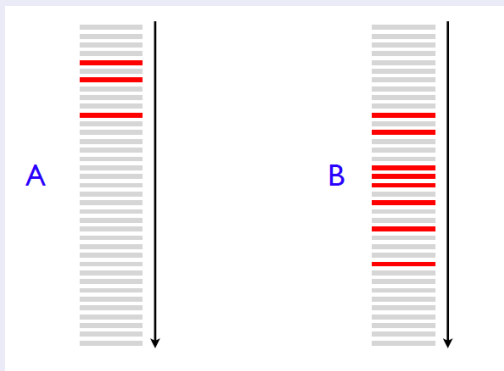
Quel est le meilleur classement ?



- Rappel au rang 10

Evaluation par bancs d'essai : métrique d'évaluation

Quel est le meilleur classement ?



- Rappel au rang 30

Evaluation par bancs d'essai

Avantages

- Si on a la collection de test : **peu coûteux**.
- Les conditions expérimentales sont fixées : mêmes requêtes et mêmes jugements de pertinence.
- Les évaluations sont reproductibles.
- L'expérimentation peut faire comprendre pourquoi et comment un système est meilleur qu'un autre.

Evaluation par bancs d'essai

Inconvénients

- La collection de test est **coûteuse** à construire.
- Les juges ne sont souvent pas les utilisateurs : **les jugements sont faits hors contexte**.
- Hypothèse forte : une pertinence indépendante de l'utilisateur et du contexte d'utilisation.

Plan

- 1 Introduction
 - Evaluation par banc d'essais
 - Evaluation par étude d'utilisateurs
 - Evaluation en ligne
- 2 Requête vs Besoin d'Information
- 3 Evaluation de résultats de recherche sans classement
- 4 Evaluation de résultats de recherche avec classement
- 5 Les collections pour l'évaluation
- 6 Présentation des résultats
- 7 Conclusion

Evaluation par étude d'utilisateurs

Principe

- Un petit ensemble d'utilisateurs avec plusieurs systèmes de RI.
- Les utilisateurs répondent à différentes tâches de recherche.
- On apprend la performance des systèmes en :
 - ▶ Observant ce que font les utilisateurs.
 - ▶ Les questionnant sur leurs actions.
 - ▶ Mesurant les succès (tâche complète, temps, ...)
 - ▶ Mesurant le succès perçu : questionnaire d'évaluation.

Evaluation par étude d'utilisateurs

Avantages

- Des données très détaillées sur le comportement des utilisateurs par rapport aux systèmes.
- Prise en compte de l'utilisateur et du contexte.

Evaluation par étude d'utilisateurs

Inconvénients

- Très coûteux.
- Difficile de généraliser les études : d'un petit ensemble d'utilisateurs à une population plus large.
- Des études en laboratoire : pas l'environnement classique d'utilisation.
- Refaire l'expérimentation dès qu'on veut tester un nouveau système.

Plan

- 1 Introduction
 - Evaluation par banc d'essais
 - Evaluation par étude d'utilisateurs
 - Evaluation en ligne
- 2 Requête vs Besoin d'Information
- 3 Evaluation de résultats de recherche sans classement
- 4 Evaluation de résultats de recherche avec classement
- 5 Les collections pour l'évaluation
- 6 Présentation des résultats
- 7 Conclusion

Evaluation en ligne

Principe

- Point de départ : un système de RI et un ensemble d'utilisateurs existants (exemple : moteurs de recherche du WEB)
- Nécessite le pourcentage de trafic d'un système par rapport à un autre.
- Comparaison à partir des logs d'interactions de l'utilisateur.
 - ▶ **clics** : substitut pour la pertinence perçue.
 - ▶ **sauts** : substitut pour la non-pertinence perçue.

Evaluation en ligne

Retour de pertinence implicite

implicit feedback in information retrieval

Advanced search

[\[PDF\] Implicit Feedback for Interactive Information Retrieval](#)
 research.microsoft.com/en-us/um/people/ryenw/papers/thesis.pdf
 File Format: PDF/Adobe Acrobat
 by RW White - 2004 - Cited by 30 - Related articles
Implicit Feedback for Interactive. **Information Retrieval**. Ryan William White.
 Department of Computing Science. Faculty of Computing Science, Mathematics and ...

[\[PDF\] Context-Sensitive Information Retrieval Using Implicit Feedback](#)
 citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.987...
 File Format: PDF/Adobe Acrobat - Quick View
 by X Shen - 2005 - Cited by 217 - Related articles
 ploit **implicit feedback** information, including previous queries and clickthrough
 information, to improve retrieval accuracy in an in- teractive **information retrieval** ...

[Context-Sensitive Information Retrieval Using Implicit Feedback](#)
 citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.987
 by X Shen - 2005 - Cited by 217 - Related articles
 CiteSeerX - Document Details (Isaac Council, Lee Giles): A major limitation ...

Show more results from psu.edu

[Relevance feedback - Wikipedia, the free encyclopedia](#)
 en.wikipedia.org/wiki/Relevance_feedback
 Relevance feedback is a feature of some **information retrieval** systems. The idea ... 1
 Explicit feedback; 2 **Implicit feedback**; 3 Blind feedback; 4 Using relevance ...

[A Search Engine that Learn from Implicit Feedback](#)
 striver.joachims.org/
 OSMOT - Learning **Retrieval** Functions from **Implicit Feedback**. ... Such observable
 behavior gives weak and noisy feedback **information** about which links the ...

click!

can we say that the
first result is more
relevant than the
second?

D'après J. Arguello

Evaluation en ligne

Retour de pertinence implicite

implicit feedback in information retrieval

Advanced search

Implicit Feedback - Under the Reading Lamp
bcao.wikidot.com/implicit-feedback
 Xuehua Shen, Bin Tan, and ChengXiang Zhai, "Context-sensitive information retrieval using **implicit feedback**," in Proceedings of the 28th annual ...

Implicit Feedback for Interactive Information Retrieval
research.microsoft.com/en-us/um/people/ryenw/papers/thesis.pdf
 File Format: PDF/Adobe Acrobat
 by RW White - 2004 - Cited by 30 - Related articles
Implicit Feedback for Interactive. **Information Retrieval**. Ryan William White.
 Department of Computing Science. Faculty of Computing Science, Mathematics and ...

Context-Sensitive Information Retrieval Using Implicit Feedback
citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.987...
 File Format: PDF/Adobe Acrobat - Quick View
 by X Shen - 2005 - Cited by 217 - Related articles
 plot **implicit feedback** information, including previous queries and clickthrough information, to improve retrieval accuracy in an in-teractive **information retrieval** ...

Context-Sensitive Information Retrieval Using Implicit Feedback
citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.987
 by X Shen - 2005 - Cited by 217 - Related articles
 CiteSeerX - Document Details (Isaac Council, Lee Giles): A major limitation ...

Show more results from psu.edu

skip!

click!

can we say that the
second result is
more relevant than
the first?


D'après J. Arguello

Evaluation en ligne

Retour de pertinence implicite

implicit feedback in information retrieval

[Images for lemurs](#) - Report Images



[Implicit Feedback for Interactive Information Retrieval](#)
research.microsoft.com/en-us/um/people/ryenw/papers/thesis.pdf
 File Format: PDF/Adobe Acrobat
 by RW White - 2004 - Cited by 30 - Related articles
Implicit Feedback for Interactive. **Information Retrieval**. Ryan William White.
 Department of Computing Science. Faculty of Computing Science, Mathematics and ...

[Context-Sensitive Information Retrieval Using Implicit Feedback](#)
citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.987...
 File Format: PDF/Adobe Acrobat - Quick View
 by X Shen - 2005 - Cited by 217 - Related articles
 exploit **implicit feedback** information, including previous queries and clickthrough
 information, to improve retrieval accuracy in an in-teractive **information retrieval** ...

[Context-Sensitive Information Retrieval Using Implicit Feedback](#)
citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.987
 by X Shen - 2005 - Cited by 217 - Related articles
 CiteSeerX - Document Details (Isaac Council, Lee Giles): A major limitation ...

[Show more results from psu.edu](#)

click!

a click is a noisy
surrogate for
relevance!


D'après J. Arguello

Evaluation en ligne

Retour de pertinence implicite

capital of honduras

[Images for lemurs](#) - Report images



[\[PDF\] Implicit Feedback for Interactive Information Retrieval](#)
research.microsoft.com/en-us/um/people/ryenw/papers/thesis.pdf
 File Format: PDF/Adobe Acrobat
 by RW White - 2004 - Cited by 30 - [Related articles](#)
Implicit Feedback for Interactive Information Retrieval, Ryan William White,
 Department of Computing Science, Faculty of Computing Science, Mathematics and ...

[\[PDF\] Context-Sensitive Information Retrieval Using Implicit Feedback](#)
citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.61.987...
 File Format: PDF/Adobe Acrobat - [Quick View](#)
 by X Shen - 2005 - Cited by 217 - [Related articles](#)
 plot **implicit feedback** information, including previous queries and clickthrough
 information, to improve retrieval accuracy in an **in-teractive information retrieval** ...

[Context-Sensitive Information Retrieval Using Implicit Feedback](#)
citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.1.61.987
 by X Shen - 2005 - Cited by 217 - [Related articles](#)
 CiteSeerX - Document Details (Isaac Councils, Lee Giles): A major limitation ...

[Show more results from psu.edu](#)

user sees the
results and
closes the
browser

D'après J. Arguello

Evaluation en ligne

Retour de pertinence implicite

capital of honduras

[Tegucigalpa Honduras](#) [maps.google.com](#)



[Tegucigalpa - Wikipedia, the free encyclopedia](#)
[en.wikipedia.org/wiki/Tegucigalpa](#)
 As **capital of Honduras**, as department head and as a municipality, the Central District seats ... For all practical purposes the **capital of Honduras** is Tegucigalpa. ...
 Etymology - History - Geography - Cityscape

[Honduras - Wikipedia, the free encyclopedia](#)
[en.wikipedia.org/wiki/Honduras](#)
 Comayagua was the **capital of Honduras** until 1880, when it was transferred to ...
 Geography of Honduras - List of cities in Honduras - Economy of Honduras - Colón
 + Show more results from wikipedia.org

[Honduras Facts and Figures, Honduras History, Political, Banking ...](#)
[www.ca-bc.com/zip_internacional/about_honduras.html](#)
 Tegucigalpa, the **capital of Honduras**, got its tongue twisting name from the ancient Nahuatl language, and translated means "silver mountain" in effect, ...

the absence of a
click is a noisy
surrogate for non-
relevance

D'après J. Arguello

Evaluation en ligne

Avantages

- Usage des systèmes réels : utilisateurs dans leur contexte d'utilisation, ne savent pas que c'est un test (pas de biais).
- L'évaluation peut se faire sur de nombreux utilisateurs.

Evaluation en ligne

Inconvénients

- Les données ne sont pas toujours disponibles : problème du *démarrage à froid*.
- Il faut bien comprendre et analyser comment les retours de pertinence implicites sont corrélés avec une expérience utilisateur positive ou négative.
- Des expériences difficiles à reproduire.

Plan

- 1 Introduction
 - Evaluation par banc d'essais
 - Evaluation par étude d'utilisateurs
 - Evaluation en ligne
- 2 Requête vs Besoin d'Information
- 3 Evaluation de résultats de recherche sans classement
- 4 Evaluation de résultats de recherche avec classement
- 5 Les collections pour l'évaluation
- 6 Présentation des résultats
- 7 Conclusion

Pertinence : besoin d'information vs requête

Exemple

- Besoin d'information : *Est ce que boire du vin rouge est plus efficace que le vin blanc concernant la réduction de risques au niveau du coeur ?*
- Requête : VIN ET ROUGE ET BLANC ET RISQUE ET COEUR ET EFFICACE

Exemple de document

Il s'est lancé dans le coeur de son discours et il a attaqué l'industrie viticole concernant sa sous-évaluation des risques du vin blanc et du vin rouge pour la conduite sous l'influence de drogues

Pertinence : besoin d'information vs requête

Exemple

- Le document est pertinent par rapport à la requête
- Le document n'est pas pertinent par rapport au besoin d'information

Conséquences

La pertinence doit être évaluée par rapport au besoin d'information et non par rapport à la requête.

Pertinence : besoin d'information vs requête

Dans l'évaluation par banc, une description du besoin d'information est associée à la requête pour juger la pertinence.

- **QUERY:** pet therapy
- **DESCRIPTION:** Relevant documents must include details of how pet- and animal-assisted therapy is or has been used. Relevant details include information about pet therapy programs, descriptions of the circumstances in which pet therapy is used, the benefits of this type of therapy, the degree of success of this therapy, and any laws or regulations governing it.

Plan

- 1 Introduction
 - Evaluation par banc d'essais
 - Evaluation par étude d'utilisateurs
 - Evaluation en ligne
- 2 Requête vs Besoin d'Information
- 3 Evaluation de résultats de recherche sans classement
- 4 Evaluation de résultats de recherche avec classement
- 5 Les collections pour l'évaluation
- 6 Présentation des résultats
- 7 Conclusion

Précision et rappel

Précision

Precision (P) est la portion des documents retrouvés qui sont pertinents

$$\text{Precision} = \frac{\#(\text{documents retrouvés pertinents})}{\#(\text{documents retrouvés})} = P(\text{pertinents}|\text{retrouvés})$$

Evaluer la capacité d'un système à renvoyer surtout des documents pertinents ou à rejeter tous les documents non-pertinents.

Rappel

Rappel (R) est la portion des documents pertinents retrouvés

$$\text{Rappel} = \frac{\#(\text{documents retrouvés pertinents})}{\#(\text{documents pertinents})} = P(\text{retrouvés}|\text{pertinents})$$

Evaluer la capacité d'un système à renvoyer tous les documents pertinents.

Précision et rappel

Précision et **Rappel** sont deux mesures couramment utilisées pour mesurer l'efficacité de la recherche

Ensemble pertinent et trouvé

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

On peut partitionner les documents en 4 ensembles :

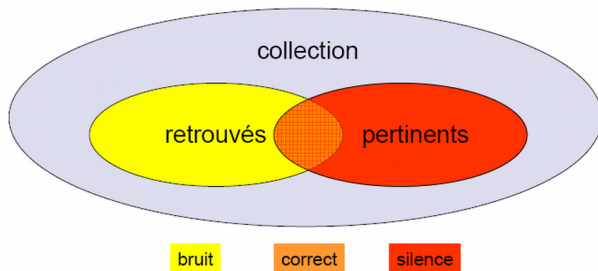
- *tp* : Vrais positifs
- *fp* : Faux positifs (le système retourne un document non pertinent)
- *fn* : Faux négatifs (le système a échoué à retourner un document pertinent)
- *tn* : Vrais négatifs

Précision et rappel

Définition

- Précision : Portion des documents trouvés qui sont pertinents.
Représente la capacité de retrouver et de bien classer les documents pour la plupart pertinents
$$\text{Precision} = \frac{tp}{(tp+fp)}$$
 (mesure dépendante du nombre de faux positifs)
- Rappel : Portion des documents pertinents qui sont retrouvés.
Représente la capacité de la recherche de trouver tous les documents pertinents du corpus
$$\text{Rappel} = \frac{tp}{(tp+fn)}$$
 (mesure dépendante du nombre de faux négatifs)

Rappel et précision



Notion de silence et de bruit.

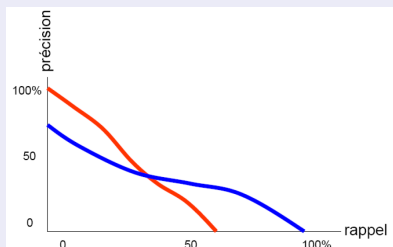
Rappel et précision

Des mesures inter-dépendantes : compromis

- La précision diminue quand le nombre de documents retournés augmente.
- Le rappel augmente quand le nombre de documents retournés augmente.
- Un système renvoyant l'ensemble des documents a un très bon rappel.
- Un système renvoyant un seul document pertinent aura une très bonne précision.

Précision et rappel

Un compromis : courbe rappel-precision



- On peut augmenter le rappel en retournant plus de documents (mais, diminution de la précision)
- Rappel = 100 % : tous les documents ont été retournés par l'algorithme de recherche
- De la même manière on peut très facilement avoir une précision haute avec un rappel très faible.

Précision et rappel

Exemples

- Systèmes pour lesquels on privilégie la précision : moteur de recherche web.
Pour toute requête, on veut que tous les documents retournés en première page soient pertinents.
- Systèmes pour lesquels on privilégie le rappel : recherche sur disque dur.
On veut que le système renvoie l'ensemble des documents pertinents en tolérant des imprécisions.

Mais souvent, on préfère utiliser une mesure unique.

Accuracy ?

Définition

- Etant donné une requête (ou un besoin), un moteur de recherche classe chaque document comme pertinent ou non pertinent
- **Accuracy** = Portion des classifications correctes
$$\text{accuracy} = \frac{(tp+tn)}{(tp+fn+fp+tn)}$$
- C'est une mesure souvent utilisée dans le monde de l'apprentissage et de la classification.
- Une mesure adaptée à la RI ?

Accuracy ?

Petit exercice

Calculer la précision, le rappel et l'accuracy sur les données suivantes :

	pertinent	non pertinent
retrouvé	18	2
non retrouvé	82	1.000.000.000

Accuracy ?

Une mesure peu adaptée

- Des données relativement biaisées : environ 99.9 % des documents sont non pertinents
- Un bon fonctionnement peut finalement amener à considérer tous les documents non pertinents pour toutes les requêtes
- Pour un RI, on veut un résultat et on a donc une tolérance sur l'erreur

Rappel et précision : deux mesures

Avantages

- Des applications différentes pour des mesures différentes.
 - ▶ recherche web : avoir le plus de documents pertinents sur la première page : précision haute.
 - ▶ recherche dans des archives : rappel haut

E measure (van Rijsbergen)

P = précision ; R = rappel

- Mesure

$$E = 1 - \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}}$$

- Permet de mettre une importance sur la précision ou le rappel (selon α)
- Moyenne pondérée de la précision et du rappel
- Avec $\alpha = 1/(\beta^2 + 1)$ et $\beta = P/R$ on a

$$E = 1 - \frac{((\beta^2+1)PR}{\beta^2 P + R}$$

- ▶ Pour $\beta = 1$, importance égale pour la précision et le rappel
- ▶ Pour $\beta > 1$ importance du rappel
- ▶ Pour $\beta < 1$ importance de la précision

F measure (Van Rijsbergen)

Une mesure souvent utilisée

- $F = 1 - E = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R}$ avec $\beta^2 = \frac{1-\alpha}{\alpha}$
- $\alpha \in [0, 1]$ et donc $\beta^2 \in [0, \infty]$.
- Bons résultats : F grand
- F1 soit $F_{\beta=1}$ ($\alpha = \frac{1}{2}$) est très populaire : poids identique sur le rappel et la précision.
- $F1 = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R} + \frac{1}{P}}$, moyenne harmonique du rappel et de la précision.
- Avec F, rappel et précision doivent être grands tous les deux pour avoir une grande mesure.
- $\beta > 1$ privilégie le rappel.
- $\beta < 1$ privilégie la précision.

Mesure du rappel et de la précision

Un travail difficile

- Le nombre total de documents pertinents n'est souvent pas disponible.
 - ▶ On peut échantillonner la collection et émettre des jugements de pertinence sur les échantillons
 - ▶ On peut appliquer différents algorithmes de recherche sur la même collection avec la même requête. On considère alors que les documents pertinents sont la somme des documents pertinents obtenus.

Plan

- 1 Introduction
 - Evaluation par banc d'essais
 - Evaluation par étude d'utilisateurs
 - Evaluation en ligne
- 2 Requête vs Besoin d'Information
- 3 Evaluation de résultats de recherche sans classement
- 4 Evaluation de résultats de recherche avec classement
- 5 Les collections pour l'évaluation
- 6 Présentation des résultats
- 7 Conclusion

Courbe rappel-précision

- Les mesures précédentes sont des mesures pour des ensembles non classés.
- L'adaptation de ces mesures à des ensembles classés est facile.
- On calcule pour chaque *préfixe* : le premier, les deux premiers,
 - ▶ L'utilisateur examine le classement de haut en bas jusqu'à ce qu'il soit satisfait.
- Precision/Rappel au rang k
- Construction d'une courbe rappel-précision.

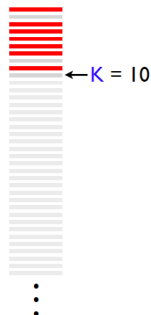
Rappel-Précision

- Précision = proportion de documents retrouvés qui sont pertinents.
- Rappel = proportion de documents pertinents qui sont retrouvés.



Rappel-Précision au rang k

- $\text{Précision}@k$ = proportion des k premiers documents retrouvés qui sont pertinents.
- $\text{Rappel}@k$ = proportion de documents pertinents qui sont dans les k premiers documents retrouvés
- L'utilisateur n'examine que les k premiers résultats.



Rappel-Précision au rang k

- Assume 20 **relevant** documents

K	P@K	R@K
1	$(1/1) = 1.0$	$(1/20) = 0.05$
2		
3		
4		
5		
6		
7		
8		
9		
10		

$K = 1$



D'après J. Arguello

Rappel-Précision au rang k

- Assume 20 **relevant** documents

K	P@K	R@K
1	$(1/1) = 1.0$	$(1/20) = 0.05$
2	$(1/2) = 0.5$	$(1/20) = 0.05$
3		
4		
5		
6		
7		
8		
9		
10		

$K = 2$



D'après J. Arguello

Rappel-Précision au rang k

Assume 20 **relevant** documents

K	$P@K$	$R@K$
1	$(1/1) = 1.0$	$(1/20) = 0.05$
2	$(1/2) = 0.5$	$(1/20) = 0.05$
3	$(2/3) = 0.67$	$(2/20) = 0.10$
4		
5		
6		
7		
8		
9		
10		

$K = 3$



D'après J. Arguello

Rappel-Précision au rang k

Assume 20 **relevant** documents

K	$P@K$	$R@K$
1	$(1/1) = 1.0$	$(1/20) = 0.05$
2	$(1/2) = 0.5$	$(1/20) = 0.05$
3	$(2/3) = 0.67$	$(2/20) = 0.10$
4	$(3/4) = 0.75$	$(3/20) = 0.15$
5		
6		
7		
8		
9		
10		

$K = 4$



D'après J. Arguello

Rappel-Précision au rang k

Assume 20 ¹relevant documents

K	$P@K$	$R@K$
1	$(1/1) = 1.0$	$(1/20) = 0.05$
2	$(1/2) = 0.5$	$(1/20) = 0.05$
3	$(2/3) = 0.67$	$(2/20) = 0.10$
4	$(3/4) = 0.75$	$(3/20) = 0.15$
5	$(4/5) = 0.80$	$(4/20) = 0.20$
6		
7		
8		
9		
10		

$K = 5$



D'après J. Arguello

Rappel-Précision au rang k

Assume 20 **relevant** documents

K	P@K	R@K
1	$(1/1) = 1.0$	$(1/20) = 0.05$
2	$(1/2) = 0.5$	$(1/20) = 0.05$
3	$(2/3) = 0.67$	$(2/20) = 0.10$
4	$(3/4) = 0.75$	$(3/20) = 0.15$
5	$(4/5) = 0.80$	$(4/20) = 0.20$
6	$(5/6) = 0.83$	$(5/20) = 0.25$
7		
8		
9		
10		

K = 6



D'après J. Arguello

Rappel-Précision au rang k

Assume 20 **relevant** documents

K	P@K	R@K
1	$(1/1) = 1.0$	$(1/20) = 0.05$
2	$(1/2) = 0.5$	$(1/20) = 0.05$
3	$(2/3) = 0.67$	$(2/20) = 0.10$
4	$(3/4) = 0.75$	$(3/20) = 0.15$
5	$(4/5) = 0.80$	$(4/20) = 0.20$
6	$(5/6) = 0.83$	$(5/20) = 0.25$
7	$(6/7) = 0.86$	$(6/20) = 0.30$
8		
9		
10		

K = 7



D'après J. Arguello

Rappel-Précision au rang k

Assume 20 **relevant** documents

K	$P@K$	$R@K$
1	$(1/1) = 1.0$	$(1/20) = 0.05$
2	$(1/2) = 0.5$	$(1/20) = 0.05$
3	$(2/3) = 0.67$	$(2/20) = 0.10$
4	$(3/4) = 0.75$	$(3/20) = 0.15$
5	$(4/5) = 0.80$	$(4/20) = 0.20$
6	$(5/6) = 0.83$	$(5/20) = 0.25$
7	$(6/7) = 0.86$	$(6/20) = 0.30$
8	$(6/8) = 0.75$	$(6/20) = 0.30$
9		
10		



D'après J. Arguello

Rappel-Précision au rang k

Assume 20 **relevant** documents

K	P@K	R@K
1	$(1/1) = 1.0$	$(1/20) = 0.05$
2	$(1/2) = 0.5$	$(1/20) = 0.05$
3	$(2/3) = 0.67$	$(2/20) = 0.10$
4	$(3/4) = 0.75$	$(3/20) = 0.15$
5	$(4/5) = 0.80$	$(4/20) = 0.20$
6	$(5/6) = 0.83$	$(5/20) = 0.25$
7	$(6/7) = 0.86$	$(6/20) = 0.30$
8	$(6/8) = 0.75$	$(6/20) = 0.30$
9	$(7/9) = 0.78$	$(7/20) = 0.35$
10		

$K = 9$



D'après J. Arguello

Rappel-Précision au rang k

Assume 20 **relevant** documents

K	$P@K$	$R@K$
1	$(1/1) = 1.0$	$(1/20) = 0.05$
2	$(1/2) = 0.5$	$(1/20) = 0.05$
3	$(2/3) = 0.67$	$(2/20) = 0.10$
4	$(3/4) = 0.75$	$(3/20) = 0.15$
5	$(4/5) = 0.80$	$(4/20) = 0.20$
6	$(5/6) = 0.83$	$(5/20) = 0.25$
7	$(6/7) = 0.86$	$(6/20) = 0.30$
8	$(6/8) = 0.75$	$(6/20) = 0.30$
9	$(7/9) = 0.78$	$(7/20) = 0.35$
10	$(7/10) = 0.70$	$(7/20) = 0.35$

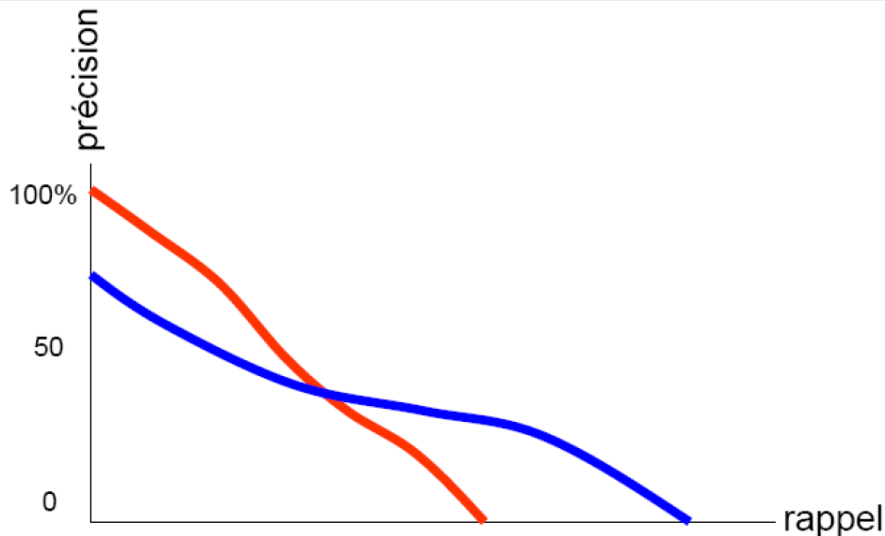
$K = 10$



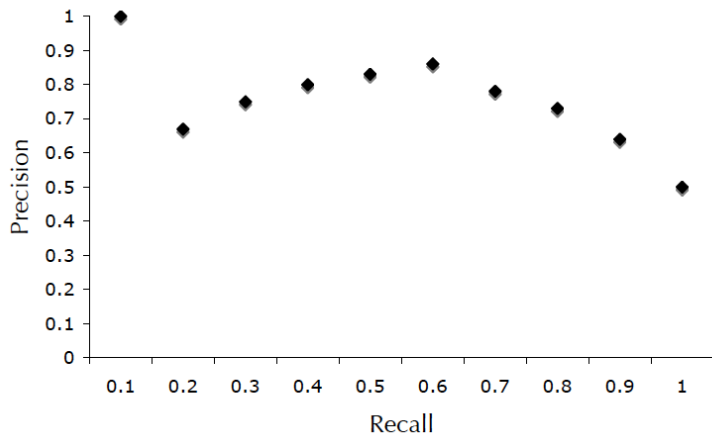
D'après J. Arguello

Précision et rappel

Courbe rappel-précision

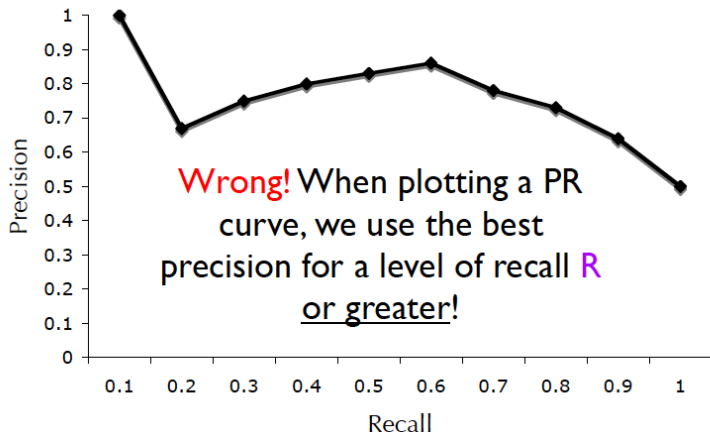


Précision et rappel : interpolation de la courbe



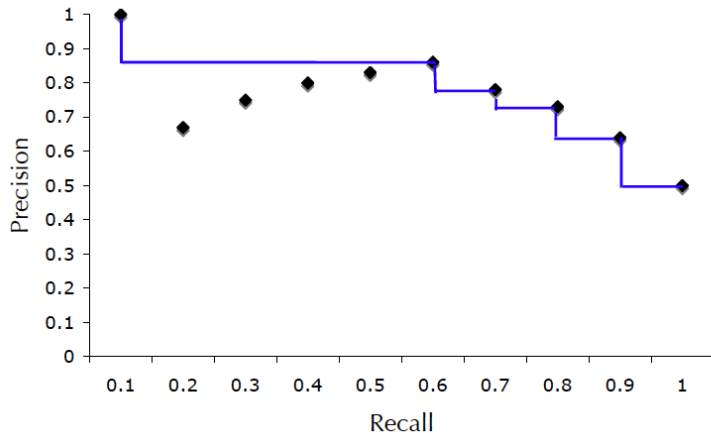
D'après J. Arguello

Précision et rappel : interpolation de la courbe



D'après J. Arguello

Précision et rappel : interpolation de la courbe



D'après J. Arguello

Précision et rappel : construction de la courbe

On suppose que l'on connaît le nombre de documents pertinents

Principe

- Pour une requête donnée, produire la liste de documents retournés avec leur classement
- Le paramétrage d'un seuil pour cette liste produit différents ensembles de documents trouvés et donc différentes mesures de rappel / précision (top k)
- Marquer chaque document dans la liste classée qui est pertinent selon un standard robuste
- Calculer la paire (precision, rappel) pour chaque position de la liste classée contenant un document pertinent

Précision et rappel : construction de la courbe

Exemple 1

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

Let total # of relevant docs = 6
Check each new recall point:

$R=1/6=0.167$; $P=1/1=1$

$R=2/6=0.333$; $P=2/2=1$

$R=3/6=0.5$; $P=3/4=0.75$

$R=4/6=0.667$; $P=4/6=0.667$

$R=5/6=0.833$; $p=5/13=0.38$

Missing one
relevant document.
Never reach
100% recall

Précision et rappel : construction de la courbe

Exemple 2

n	doc #	relevant
1	588	x
2	576	
3	589	x
4	342	
5	590	x
6	717	
7	984	
8	772	x
9	321	x
10	498	
11	113	
12	628	
13	772	
14	592	x

Let total # of relevant docs = 6
Check each new recall point:

$R=1/6=0.167$; $P=1/1=1$

$R=2/6=0.333$; $P=2/3=0.667$

$R=3/6=0.5$; $P=3/5=0.6$

$R=4/6=0.667$; $P=4/8=0.5$

$R=5/6=0.833$; $P=5/9=0.556$

$R=6/6=1.0$; $p=6/14=0.429$

Précision et rappel : interpolation de la courbe

Principe

- Interpolation d'une valeur de précision par chaque niveau de rappel :
 - ▶ $r_j \in \{0.0, 0.1, \dots, 1\}$
- La précision interpolée au niveau j est la précision maximum connue pour les points futurs, $r_i > r_j$:
 - ▶ $P(r_j) = \max_{r_i \geq r_j} P(r_i)$
- Interprétation de cette interpolation : pourcentage de documents pertinents qu'un observateur observera s'il vaut atteindre un taux de rappel au moins égal à r_j

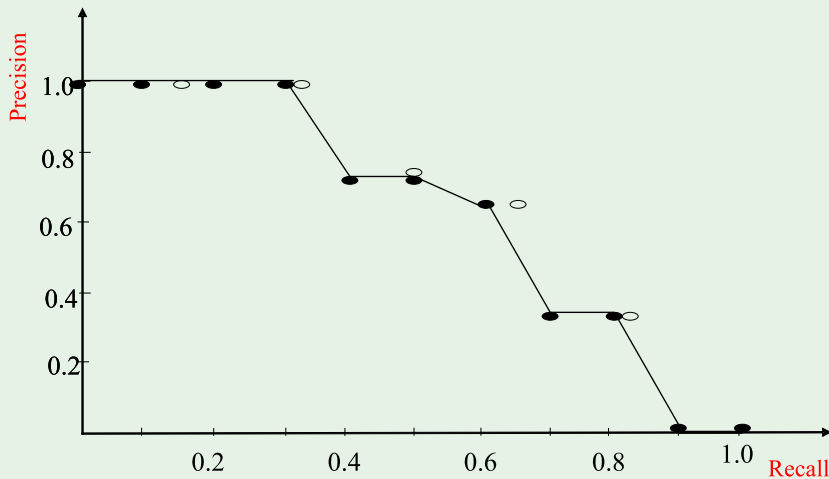
Précision et rappel : interpolation de la courbe

Interpolation pour l'exemple 1

Rappel	Precision
0,0	1
0,1	1
0,2	1
0,3	1
0,4	0,75
0,5	0,75
0,6	0,667
0,7	0,38
0,8	0,38
0,9	0
1	0

Précision et rappel : interpolation de la courbe

Courbe pour l'exemple 1



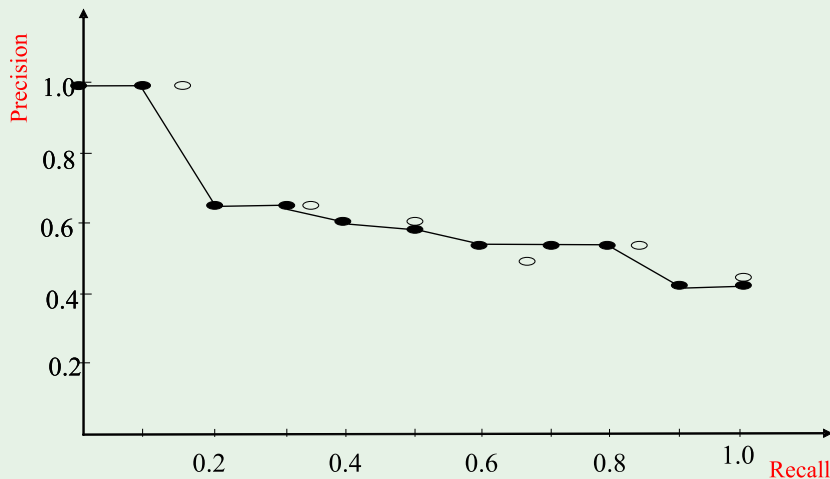
Précision et rappel : interpolation de la courbe

Interpolation pour l'exemple 2

Rappel	Precision
0,0	1
0,1	1
0,2	0,667
0,3	0,667
0,4	0,6
0,5	0,6
0,6	0,556
0,7	0,556
0,8	0,556
0,9	0,429
1	0,429

Précision et rappel : interpolation de la courbe

Courbe pour l'exemple 2

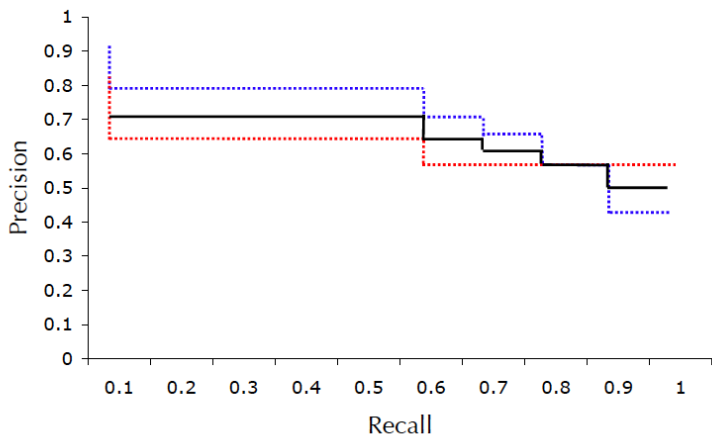


Courbe moyenne à 11 points

- Moyenne de la performance sur un grand nombre de requêtes
- Pour chaque besoin d'information (requête), la précision est interpolée pour les 11 niveaux de rappels.
- Pour chaque niveau, calcul de la moyenne arithmétique de la précision sur l'ensemble des requêtes
- Tracé de la courbe moyenne pour évaluer la performance du système sur un corpus
- La comparaison de systèmes revient à comparer les courbes : la courbe la plus proche du coin droit indique la meilleure performance

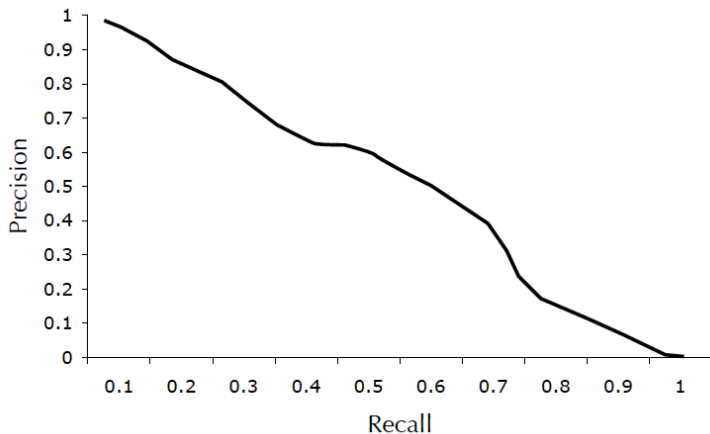
Une mesure utilisée pour les 8 premiers challenges *Recherche Ad Hoc* du challenge TREC.

Précision et rappel : interpolation de la courbe



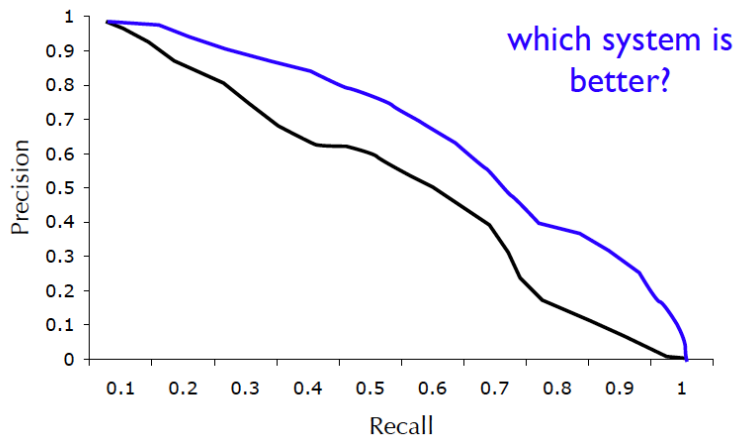
D'après J. Arguello

Précision et rappel : interpolation de la courbe



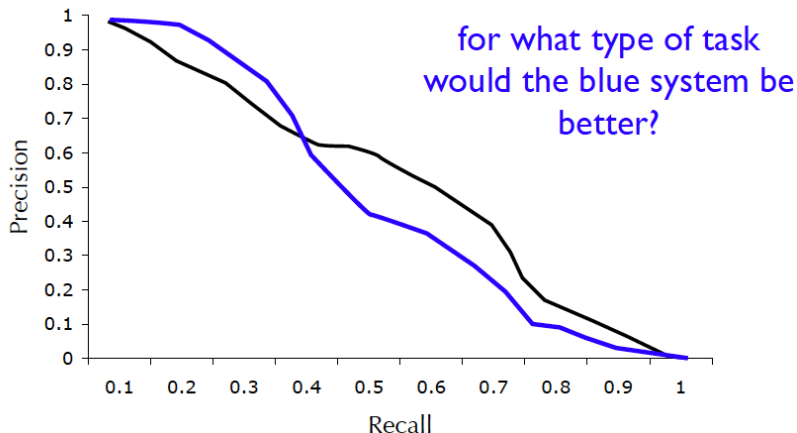
D'après J. Arguello

Précision et rappel : interpolation de la courbe



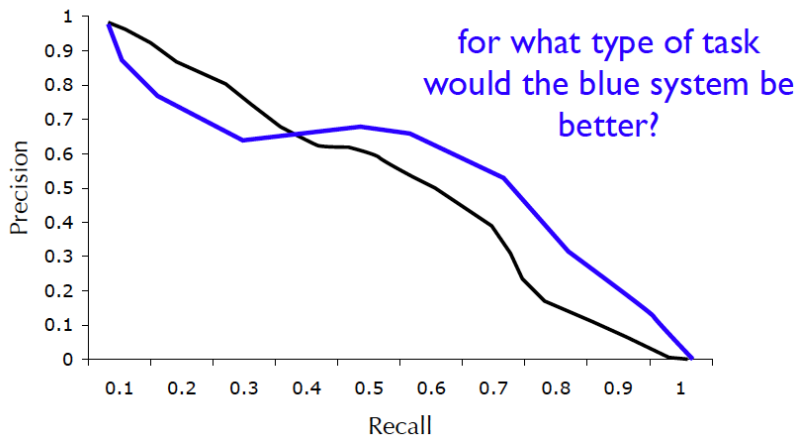
D'après J. Arguello

Précision et rappel : interpolation de la courbe



D'après J. Arguello

Précision et rappel : interpolation de la courbe



D'après J. Arguello

Mesures uniques

- Souvent utile d'avoir un nombre unique pour caractériser la performance
 - ▶ Précision pour n documents retrouvés
 - ▶ Précision moyenne : moyenne de la précision pour chaque document pertinent retrouvé. La précision d'un document pertinent non trouvé est 0
 - ▶ R-précision : précision pour ($\#$ documents pertinents) retrouvés, i.e. précision à la R ème position dans la liste classée pour une requête qui a R documents pertinents

Mesures uniques

R-précision

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

$R = \# \text{ of relevant docs} = 6$

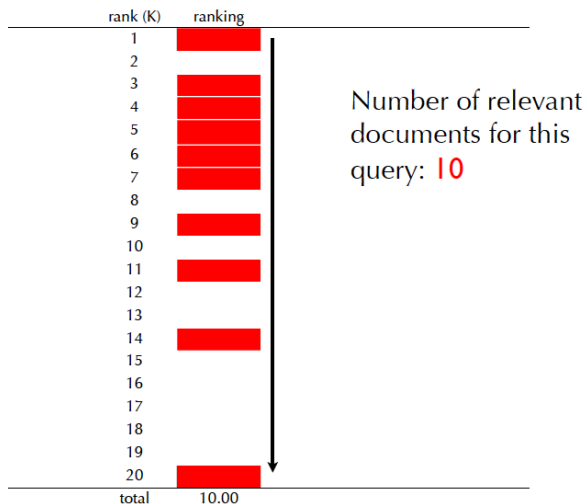
$R\text{-Precision} = 4/6 = 0.67$

Mesures uniques : précision moyenne

Principe

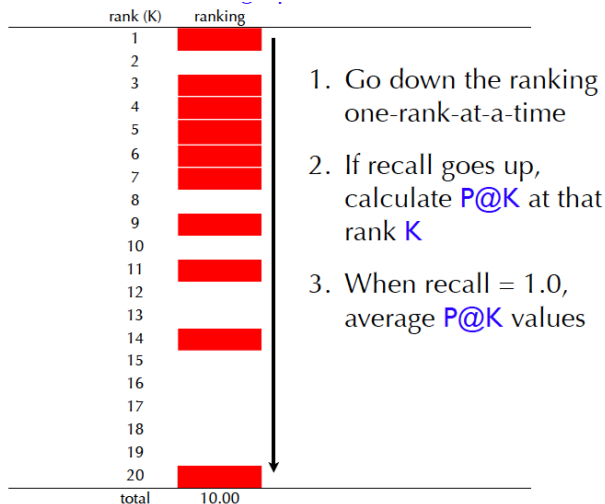
- 1 Parcourir le classement rang par rang.
- 2 Si le document au rang k est pertinent, calculer $\text{Precision}@k$.
- 3 Faire la moyenne de toutes les valeurs $\text{Precision}@k$ quand le rappel vaut 1.

Mesures uniques : précision moyenne



D'après J. Arguello

Mesures uniques : précision moyenne



D'après J. Arguello

Mesures uniques : précision moyenne

Précision moyenne

Moyenne des valeurs de précision des documents pertinents par rapport à q dans la liste ordonnée des réponses/

$$AveP(q) = \frac{1}{n_+^q} \sum_{k=1}^N R_{d_k, q} \times P@ (k)(q)$$

avec $n_+^q = \sum_{i=1}^N R_{d_i, q}$ le nombre total de documents pertinents par rapport à q

Souvent on prend la précision interpolée :

$$AveP(q) = \frac{1}{11} \sum_{r_j \in [0, 0.1, 0.2, \dots, 1.0]} P(r_j)$$

Mesures uniques : précision moyenne

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.10	0.50
3		0.20	0.67
4		0.30	0.75
5		0.40	0.80
6		0.50	0.83
7		0.60	0.86
8		0.60	0.75
9		0.70	0.78
10		0.70	0.70
11		0.80	0.73
12		0.80	0.67
13		0.80	0.62
14		0.90	0.64
15		0.90	0.60
16		0.90	0.56
17		0.90	0.53
18		0.90	0.50
19		0.90	0.47
20		1.00	0.50
total	10.00	average-precision	0.76

D'après J. Arguello

Mesures uniques : précision moyenne

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.20	1.00
3		0.30	1.00
4		0.40	1.00
5		0.50	1.00
6		0.60	1.00
7		0.70	1.00
8		0.80	1.00
9		0.90	1.00
10		1.00	1.00
11		1.00	0.91
12		1.00	0.83
13		1.00	0.77
14		1.00	0.71
15		1.00	0.67
16		1.00	0.63
17		1.00	0.59
18		1.00	0.56
19		1.00	0.53
20		1.00	0.50
total	10.00	average-precision	1.00

D'après J. Arguello

Mesures uniques : précision moyenne

rank (K)	ranking	R@K	P@K
1		0.00	0.00
2		0.00	0.00
3		0.00	0.00
4		0.00	0.00
5		0.00	0.00
6		0.00	0.00
7		0.00	0.00
8		0.00	0.00
9		0.00	0.00
10		0.00	0.00
11		0.10	0.09
12		0.20	0.17
13		0.30	0.23
14		0.40	0.29
15		0.50	0.33
16		0.60	0.38
17		0.70	0.41
18		0.80	0.44
19		0.90	0.47
20		1.00	0.50
total	10.00	average-precision	0.33

D'après J. Arguello

Mesures uniques : précision moyenne

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.10	0.50
3		0.20	0.67
4		0.30	0.75
5		0.40	0.80
6		0.50	0.83
7		0.60	0.86
8		0.60	0.75
9		0.70	0.78
10		0.70	0.70
11		0.80	0.73
12		0.80	0.67
13		0.80	0.62
14		0.90	0.64
15		0.90	0.60
16		0.90	0.56
17		0.90	0.53
18		0.90	0.50
19		0.90	0.47
20		1.00	0.50
total	10.00	average-precision	0.76

D'après J. Arguello

Mesures uniques : précision moyenne

swapped
ranks 2 and 3

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.20	1.00
3		0.20	0.67
4		0.30	0.75
5		0.40	0.80
6		0.50	0.83
7		0.60	0.86
8		0.60	0.75
9		0.70	0.78
10		0.70	0.70
11		0.80	0.73
12		0.80	0.67
13		0.80	0.62
14		0.90	0.64
15		0.90	0.60
16		0.90	0.56
17		0.90	0.53
18		0.90	0.50
19		0.90	0.47
20		1.00	0.50
total	10.00	average-precision	0.79

D'après J. Arguello

Mesures uniques : précision moyenne

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.10	0.50
3		0.20	0.67
4		0.30	0.75
5		0.40	0.80
6		0.50	0.83
7		0.60	0.86
8		0.60	0.75
9		0.70	0.78
10		0.70	0.70
11		0.80	0.73
12		0.80	0.67
13		0.80	0.62
14		0.90	0.64
15		0.90	0.60
16		0.90	0.56
17		0.90	0.53
18		0.90	0.50
19		0.90	0.47
20		1.00	0.50
total	10.00	average-precision	0.76

D'après J. Arguello

Mesures uniques : précision moyenne

swapped ranks
8 and 9

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.10	0.50
3		0.20	0.67
4		0.30	0.75
5		0.40	0.80
6		0.50	0.83
7		0.60	0.86
8		0.70	0.88
9		0.70	0.78
10		0.70	0.70
11		0.80	0.73
12		0.80	0.67
13		0.80	0.62
14		0.90	0.64
15		0.90	0.60
16		0.90	0.56
17		0.90	0.53
18		0.90	0.50
19		0.90	0.47
20		1.00	0.50
total	10.00	average-precision	0.77

D'après J. Arguello

Mesures uniques : précision moyenne

Avantages

- Pas besoin de choisir k .
- Prend en compte à la fois le rappel et la précision.
- Des erreurs en début sont plus influentes.
- Des erreurs à la fin comptent aussi

Inconvénients

- Pas facile à interpréter.

Mean Average Precision (MAP)

Mesure très standard dans la communauté TREC.

- On peut mesurer l'efficacité d'une liste ordonnée par rapport à une requête
- On cherche la performance pour une requête arbitraire
- Jusqu'à maintenant (sauf pour la courbe), nous n'avons considéré qu'une seule requête.
 - ▶ Moyenne de la précision moyenne sur différentes requêtes pour trouver l'efficacité moyenne

Mean Average Precision (MAP)

- On cherche la précision moyenne pour chaque requête (Average Precision)
- On calcule la précision moyenne sur toutes les requêtes (Mean Average Precision)

Définition

Soit $\mathcal{Q} = \{q_1, \dots, q_{|\mathcal{Q}|}\}$, un ensemble de requêtes

$$MAP(\mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} AveP(q_j)$$

$$MAP(\mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} \frac{1}{n_+^{q_j}} \sum_{k=1}^N R_{d_k, q_j} \times P@k(q_j)$$

Gain cumulatif réduit (Discounted Cumulative Gain)(DCG)

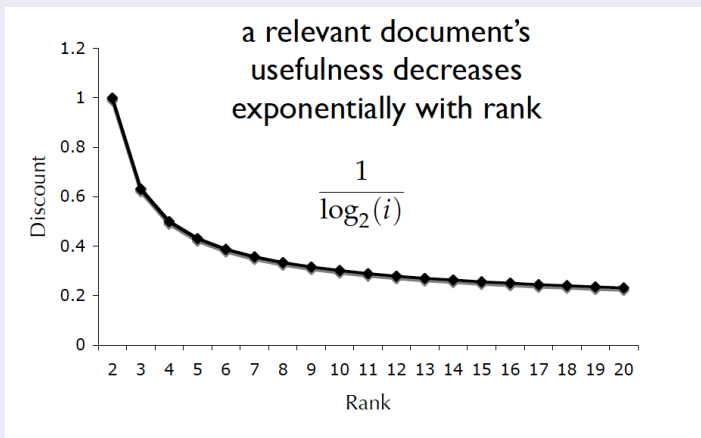
- Evaluation des SRI pour lesquels les jugements de pertinence ne sont plus binaires mais à valeurs dans un sous-ensemble de \mathbb{N} (notion d'échelle de pertinence).
- Évalue l'utilité d'un document en fonction de son score de pertinence et somme ce gain pour les k premiers documents retournés en prenant en compte leur rang.

$$DCG@k(q) = \sum_{r=1}^k \frac{rel(d_r, q)}{\log_2(\max(r, 2))}$$

avec $rel(d_r, q)$ le score de pertinence du document au rang r en fonction de la requête q .

Gain cumulatif réduit (Discounted Cumulative Gain)(DCG)

Principe



Source : J. Arguello

Gain cumulatif réduit (Discounted Cumulative Gain)(DCG)

Principe

rank (i)	REL_i	
1	4	This is given!
2	3	
3	4	
4	2	the result at rank 1 is perfect
5	0	the result at rank 2 is excellent
6	0	the result at rank 3 is perfect
7	0	...
8	1	the result at rank 10 is bad
9	1	
10	0	

Source : J. Arguello

Gain cumulatif réduit (Discounted Cumulative Gain)(DCG)

Principe

rank (i)	REL_i	discount factor	
1	4	1.00	Each rank is associated with a discount factor
2	3	1.00	
3	4	0.63	
4	2	0.50	
5	0	0.43	$\frac{1}{\log_2(\max(i, 2))}$ rank 1 is a special case!
6	0	0.39	
7	0	0.36	
8	1	0.33	
9	1	0.32	
10	0	0.30	

Source : J. Arguello

Gain cumulatif réduit (Discounted Cumulative Gain)(DCG)

Principe

rank (i)	REL _i	discount factor	gain
1	4	1.00	4.00
2	3	1.00	3.00
3	4	0.63	2.52
4	2	0.50	1.00
5	0	0.43	0.00
6	0	0.39	0.00
7	0	0.36	0.00
8	1	0.33	0.33
9	1	0.32	0.32
10	0	0.30	0.00

multiply REL_i
by the
discount
factor
associated
with the
rank!

Source : J. Arguello

Gain cumulatif réduit (Discounted Cumulative Gain)(DCG)

Principe

rank (i)	REL_i	discount factor	gain	DCG_i
1	4	1.00	4.00	4.00
2	3	1.00	3.00	7.00
3	4	0.63	2.52	9.52
4	2	0.50	1.00	10.52
5	0	0.43	0.00	10.52
6	0	0.39	0.00	10.52
7	0	0.36	0.00	10.52
8	1	0.33	0.33	10.86
9	1	0.32	0.32	11.17
10	0	0.30	0.00	11.17

Source : J. Arguello

Gain cumulatif réduit normalisé (Normalized Discounted Cumulative Gain)(NDCG)

- Normalisation de la mesure précédente en divisant le gain cumulatif par celui obtenu avec un système idéal parfait qui placerait en tête les documents les plus pertinents (ceux de plus grand score de pertinence)

$$NDCG@k(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_{kj} \sum_{r=1}^k \frac{rel(d_r, q)}{\log_2(max(r, 2))}$$

avec $rel(d_r, q)$ le score de pertinence du document au rang r en fonction de la requête q et Z_{kj} facteur de normalisation.

Plan

- 1 Introduction
 - Evaluation par banc d'essais
 - Evaluation par étude d'utilisateurs
 - Evaluation en ligne
- 2 Requête vs Besoin d'Information
- 3 Evaluation de résultats de recherche sans classement
- 4 Evaluation de résultats de recherche avec classement
- 5 Les collections pour l'évaluation**
- 6 Présentation des résultats
- 7 Conclusion

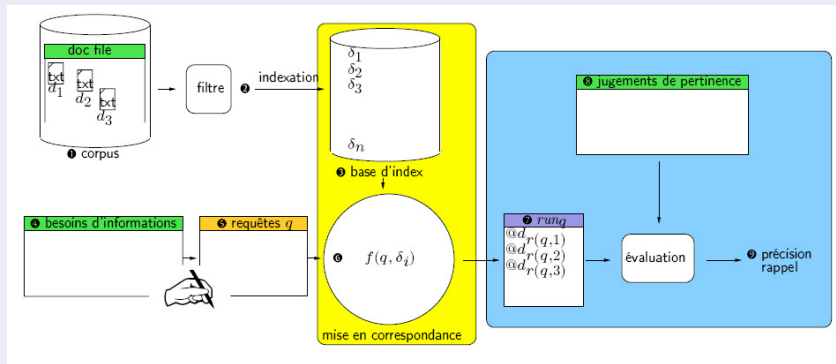
Points importants

Besoins pour un banc d'essais

- Un corpus de documents (représentatifs des documents que l'on cherche)
- Un corpus de besoins d'information (souvent sous forme de requêtes et représentatifs de la réalité)
- Des jugements de pertinence (qui doivent être consistents, i.e. sans contradiction (mesure Kappa))

Evaluation par benchmarks

Principe



Evaluation par benchmarks

Principe

- 1 Documents originaux
- 2 Fichiers de la collection
- 3 Index des documents
- 4 Les besoins d'information
- 5 Les requêtes
- 6 L'index des requêtes
- 7 Les listes de réponses retournées par le système de RI
- 8 Les ensembles de documents jugés pertinents
- 9 L'évaluation précision-rappel

Evaluation par benchmarks

Collection **Cranfield**

- Benchmark pionnier permettant la mesure quantitative de l'efficacité des RI
- 1950, Royaume-Uni
- 1398 résumés d'articles de journaux en aerodynamisme, 225 requêtes, jugements de pertinence exhaustifs pour chaque paire (document, requête)

Cette méthode d'évaluation est aussi appelée méthode de Cranfield

Evaluation par benchmarks

Collection **SMART**

- Premières expériences sur cette petite collection

Collection Name	Number Of Documents	Number Of Queries	Raw Size (Mbytes)
CACM	3,204	64	1.5
CISI	1,460	112	1.3
CRAN	1,400	225	1.6
MED	1,033	30	1.1
TIME	425	83	1.5

Evaluation par benchmarks

Les grandes campagnes d'évaluation

- TREC : Text Retrieval Conference
<http://trec.nist.gov/>
- CLEF : Cross Language Evaluation Forum
<http://www.clef-initiative.eu/>
- NTCIR : NII Testbeds and Community for Information access Research
<http://research.nii.ac.jp/ntcir/index-en.html>
- INEX : Initiative for the Evaluation of XML retrieval
<http://inex.mmci.uni-saarland.de/>
- FIRE : Forum for Information Retrieval Evaluation
<http://fire.irsu.res.in/fire/home>

Evaluation par benchmarks

Collection **TREC**(Text Retrieval Conference)

<http://trec.nist.gov/>

- Campagne d'évaluation organisée par le U.S. National Institute of Standards and Technology (NIST) depuis 1992.
- Actuellement, ensemble de différents benchmarks et différents tracks (web track, Chemical IR track, entity track, Legal Track, Million Query Track, Relevance Feedback Track et Web Track)
- Le plus connu : TREC Ad Hoc utilisé pour les 8 premières campagnes d'évaluation
- 1.89 million de documents, principalement des articles de presse et 450 besoins d'information (*topics*)
- Pas de jugements de pertinence exhaustifs mais uniquement sur les k documents retournés par un ensemble de systèmes de RI

Evaluation par benchmarks

Autres corpus

<http://www.hlt-evaluation.org/spip.php?article144>

- GOV2 (http://ir.dcs.gla.ac.uk/test_collections/gov2-summary.htm)
 - ▶ TREC NIST Collection
 - ▶ 25 millions de pages web (sites web de domaines .gov en 2004) : 0.42 To...
- ClueWeb09
 - ▶ TREC NIST Collection
 - ▶ 1 billion de pages web dans 10 languages (collectés en 2009) : 25 To (sans compression)
 - ▶ Une des plus grandes collections disponibles mais encore inférieure aux moteurs de recherche actuel (Google,...)
 - ▶ 4.780.950.613 URLs et 454.075.638 liens sortants.
- ClueWeb12

Evaluation par benchmarks

Autres corpus

- NTCIR : NII Collections for IR Systems (<http://research.nii.ac.jp/ntcir/>)
 - ▶ Focalisé sur la langue asiatique et le multi-linguisme
- CLEF : Cross Language Evaluation Forum (<http://www.clef-campaign.org/>)
 - ▶ Focalisé sur les langues européenne et le multi-linguisme
- Reuters : <http://trec.nist.gov/data/reuters/reuters.html>
- Newsgroup, Blog : <http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG/>

Validité des jugements de pertinence

Points importants

- Les jugements de pertinences ne sont utilisables que si ils sont consistants entre eux.
- Dans le cas contraire, les expériences ne peuvent pas être reproductibles.
- Comment mesurer cette consistance (ou accord entre les jugements) ?
 - ▶ Mesure Kappa.

Validité des jugements de pertinence

Mesure Kappa

- Mesure l'accord ou le désaccord entre les jugements.
- Mesure l'accord entre observateurs lors d'un codage qualitatif en catégories.

$$\kappa = \frac{P(A) - P(E)}{(1 - P(E))}$$

- $P(A)$: accord relatif entre juges (nombre (proportion) de fois que les juges sont d'accord)
- $P(E)$: probabilité d'un accord aléatoire (nombre (proportion) de fois ou l'accord est aléatoire)

Validité des jugements de pertinence

Mesure Kappa

- Des valeurs de κ dans l'intervalle $[0.67, 1]$ sont acceptables.
- Si la valeur est plus petite, il faut revoir le processus d'obtention des jugements de pertinence.
- Calcul de $P(E)$: 0.5 (cas binaire) ou statistiques marginales (opposées de la distribution conjointe, on ne prend pas en compte les autres variables).

$$\text{e.g. } P(E) = P(\text{nonrelevant})^2 + P(\text{relevant})^2$$

Validité des jugements de pertinence

Mesure Kappa : exemple de calcul

		Judge 2 Relevance		
		Yes	No	Total
Judge 1 Relevance	Yes	300	20	320
	No	10	70	80
	Total	310	90	400

Calculating the kappa statistic

		Judge 2 Relevance		
		Yes	No	Total
Judge 1 Relevance	Yes	300	20	320
	No	10	70	80
	Total	310	90	400

- $P(A) = (300 + 70)/400 = 370/400 = 0.925$
- $P(\text{nonrelevant}) = (80 + 90)/(400 + 400) = 170/800 = 0.2125$
- $P(\text{relevant}) = (320 + 310)/(400 + 400) = 630/800 = 0.7875$
- Probabilité que deux juges soient d'accord par chance
 $P(E) = P(\text{nonrelevant})^2 + P(\text{relevant})^2 = 0.2125^2 + 0.7875^2 = 0.665$
- $\kappa = (P(A) - P(E))/(1 - P(E)) = (0.925 - 0.665)/(1 - 0.665) = 0.776$

Techniques du Pooling

Problème du rappel dans de grandes collections

- Le rappel impose en théorie de connaître tous les documents pertinents.
- Impossible en pratique d'avoir des jugements de pertinence exhaustifs.
- \Rightarrow : technique du pooling.

Pooling

- Fusion intelligente des résultats.
- Les k ($k > 100$) premiers documents produits par des systèmes de RI sont fusionnés.
- Seuls ces documents sont jugés par les experts humains.
- Les documents non jugés sont considérés comme non pertinents.
- Le calcul du rappel se fait comme si tout avait été jugé.

Techniques du Pooling

Selon (Zobel 98), seulement 50 à 70 % des documents pertinents seraient retrouvés par cette méthode.

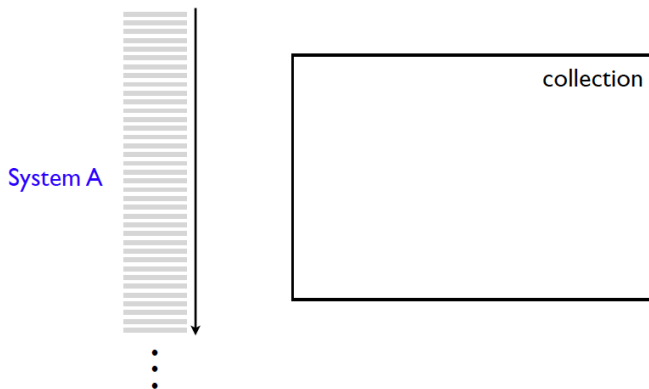
Biais du pooling

- Le rappel est sur-évalué
- La précision est sous-évaluée

Mais

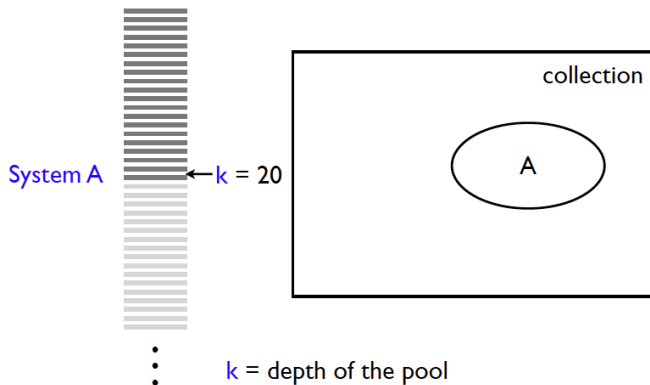
- Biais faible si il y a suffisamment de requêtes et de systèmes.
- L'évaluation relative, c'est à dire la comparaison entre systèmes reste valable
- Finalement, pas trop le choix.

Techniques du Pooling



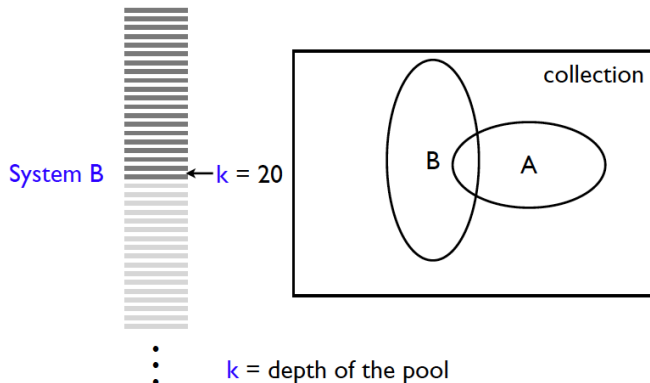
D'après J. Arguello

Techniques du Pooling



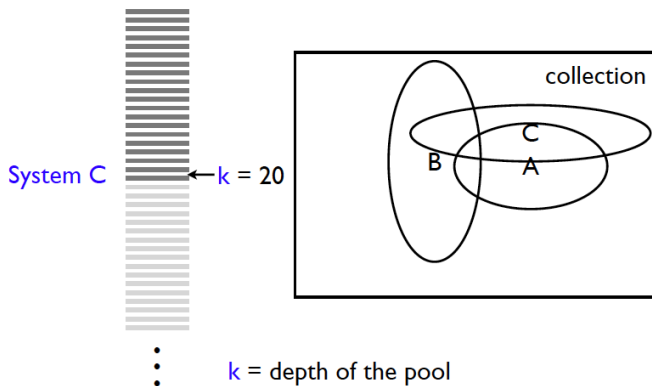
D'après J. Arguello

Techniques du Pooling



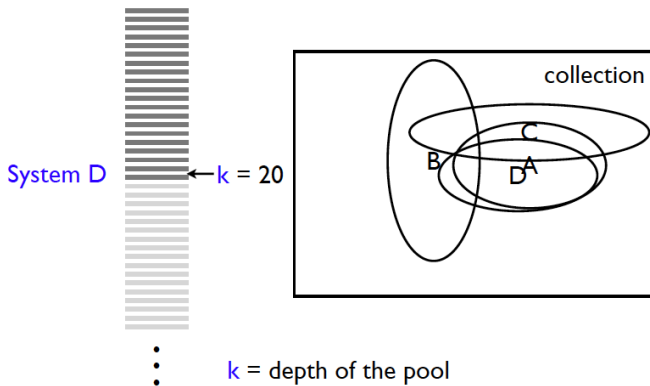
D'après J. Arguello

Techniques du Pooling



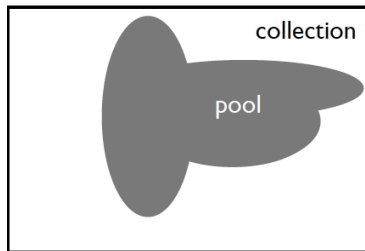
D'après J. Arguello

Techniques du Pooling



D'après J. Arguello

Techniques du Pooling



D'après J. Arguello

Plan

- 1 Introduction
 - Evaluation par banc d'essais
 - Evaluation par étude d'utilisateurs
 - Evaluation en ligne
- 2 Requête vs Besoin d'Information
- 3 Evaluation de résultats de recherche sans classement
- 4 Evaluation de résultats de recherche avec classement
- 5 Les collections pour l'évaluation
- 6 Présentation des résultats
- 7 Conclusion

Comment présenter la liste de résultats à l'utilisateur ?

- Le plus souvent comme une liste de n documents.
- Cette liste doit fournir une information à l'utilisateur car l'utilisateur ne pourra certainement parcourir tous les éléments de la liste.
- Les documents retournés sont souvent accompagnés d'une description résumée (*snippet*)
- Cette description doit permettre à l'utilisateur de décider de la pertinence du document par rapport à son besoin d'information.

Description du document

- Très souvent : titre du document, url, ensemble de métadonnées.
- Un résumé.
- Comment construire ce résumé ?

Création de résumés

Deux approches

- **Résumé statique** : toujours le même, quelque soit la requête.
- **Résumé dynamique** : dépend de la requête en tentant d'expliquer pourquoi le document a été retourné étant donnée la requête.

Résumés statiques

- Souvent un sous-ensemble du document.
- Cas le plus simple : les 50 premiers mots du document par exemple.
- Techniques plus sophistiquées : extraire de chaque document un ensemble de phrases clés.
 - ▶ Utilisation de techniques issues du traitement du langage naturel.
 - ▶ Utilisation de techniques d'apprentissage statistique.

Résumés dynamiques

- Présentent une ou plusieurs *fenêtres* du document qui permettent de présenter les parties du document qui sont utiles à évaluer la pertinence du document par rapport au besoin d'information.
- Ces fenêtres contiennent souvent 1 ou plusieurs termes de la requête (KWIC snippets : keyword in context snippet).
- Les utilisateurs préfèrent souvent des résumés dans lesquels les termes de la requête apparaissent dans une phrase.
- Exemple google.

Plan

- 1 Introduction
 - Evaluation par banc d'essais
 - Evaluation par étude d'utilisateurs
 - Evaluation en ligne
- 2 Requête vs Besoin d'Information
- 3 Evaluation de résultats de recherche sans classement
- 4 Evaluation de résultats de recherche avec classement
- 5 Les collections pour l'évaluation
- 6 Présentation des résultats
- 7 Conclusion

Points importants

- Qu'est ce qu'une bonne mesure ?
 - ▶ En rapport avec un modèle d'utilisateur et la notion de pertinence.
- Stabilité de la mesure
- Une évaluation par campagnes.