RECHERCHE D'INFORMATION & TRAITEMENT AUTOMATIQUE DU LANGAGE

Cours 3 : RI - Evaluation en RI

2022-23

Benjamin Piwowarski / Laure Soulier

ISIR INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

MLIA Machine Learning & Deep Learning for Information Access

Enjeux de l'évaluation en RI

- Quel modèle de RI est le plus efficace ?
- Efficace ? Problème difficile, pas de mesure absolue
  - Critères de qualité d'un système de RI
  - Facilité d'utilisation du système
  - Coût accès/stockage
  - Présentation des résultats
  - Efficacité de la recherche
  - Possibilités de formuler des requêtes riches
- Nombreuses mesures donnent des renseignements partiels sur le comportement du système

Evaluation Cranfield

Vers un protocole d'évaluation standardisé

- Objectif : évaluer la capacité d'un système à retourner des documents pertinents
- De quoi a-t-on besoin ?

Paradigme de Cranfield - première expérimentation "laboratoire" en RI

Evaluation basées sur des collections de test composées de :
- Corpus de documents
- Requêtes
- Jugements de pertinence

- Avantages
  - Peu couteux
  - Facilite les analyses d'erreurs
  - Répétables
- Inconvénients
  - Jugements de pertinence peuvent être incomplets
  - Quelles hypothèses pour la pertinence ?

Evaluation Cranfield : Exemples de corpus de test

- De nombreuses collections de test ont été développés par la communauté scientifique
  - Cranfield fin des années 50
  - TREC (NIST)
  - CLEF
  - NTCIR (Japon et autres langues asiatiques, cross language evaluation)
  - ...

Evaluation Cranfield : Exemples de corpus de test

- Ad hoc Test Collections
- Web Test Collections
- Blog Track
- Chemical IR Track
- Clinical Decision Support Track
- Common Core Track
- Confusion Track
- Contextual Suggestion Track
- Interactive Track
- Knowledge Base Acceleration Track
- Legal Track
- Medical Track
- Microblog Track
- Million Query Track
- Novelty Track
- Query Track
- Question Answering Track
- Precision Medicine Track
- Real-time Summarization Track
- Relevance Feedback Track
- Robust Track
- Session Track
- SPAM Track
- Spoken Document Retrieval Track
- Tasks Track
- Temporal Summarization Track
- Terabyte Track
- Web Track

Evaluation Cranfield : Documents

SCIENCES SORBONNE UNIVERSITÉ

```
<DOC>
<DOCNO> GPXX-0002 </DOCNO>
<TITLE> ceci est un titre </TITLE>
<DATE> 2006-09-04 </DATE>
<TEXT> cecie est le contenu blablabla blablabla blablabla .</TEXT>
</DOC>
```

Evaluation Cranfield : Requêtes

- Souvent les requêtes (mots clés) sont associés à une description plus complète (phrases) du besoin d'information

<top>

<num> Number : 501
<title> deduction and induction in English ?

<desc> Description :
What is the difference between deduction and induction in the process of reasoning ?

<narr> Narrative :
A relevant document will contrast inductive and deductive reasoning.
A document that discusses only one or the other is not relevant.

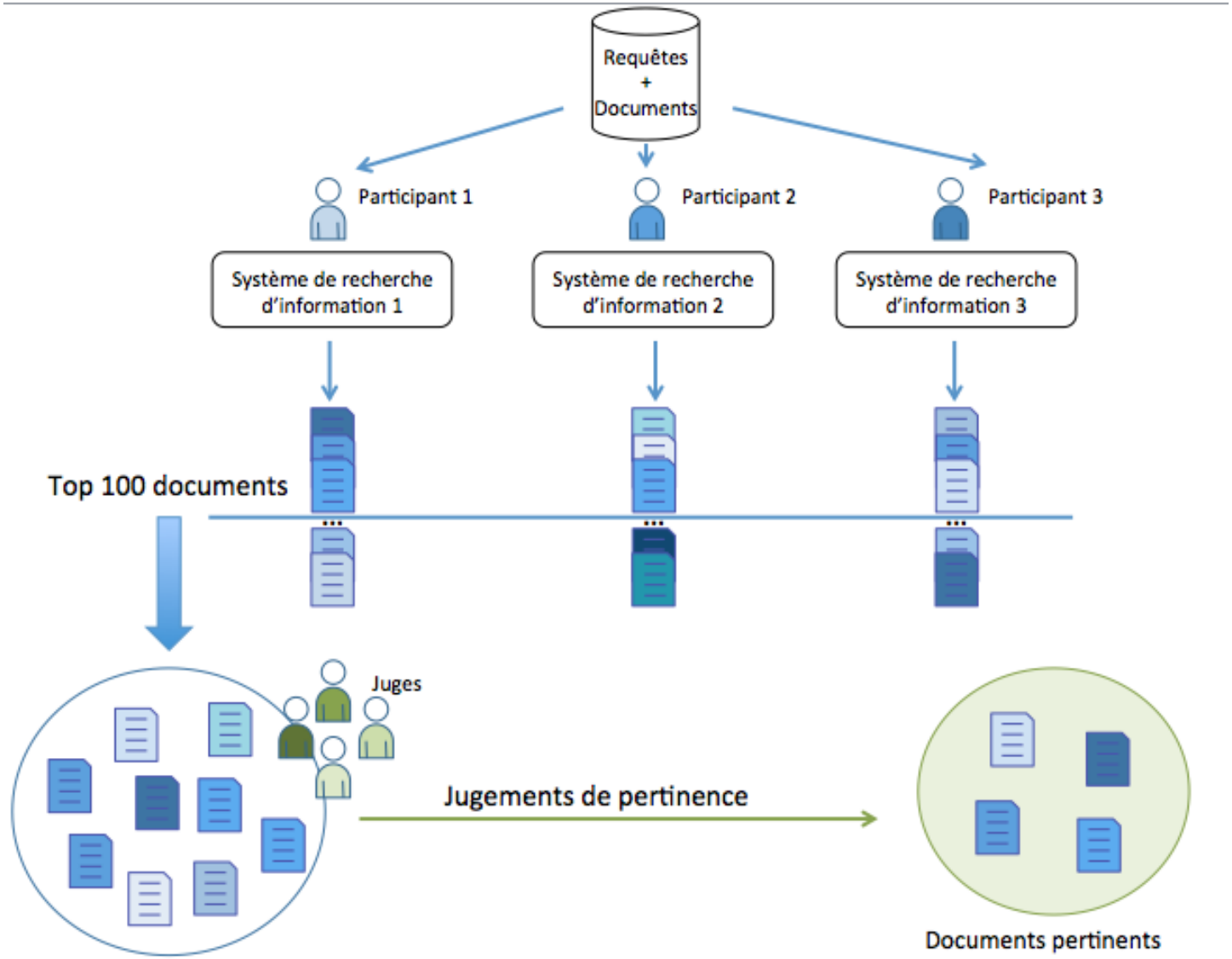
</top>

Evaluation Cranfield : Jugements de pertinence

SCIENCES SORBONNE UNIVERSITÉ

- Ensemble de jugements de pertinence pour chaque paire (document, requête)
  - Ces jugements peuvent être binaires ou être donnés sous forme d'un score, typiquement 0, 1, 2, 3, 4, 5
  - Ils sont formulés en fonction du besoin d'information

Evaluation Cranfield : Jugements de pertinence

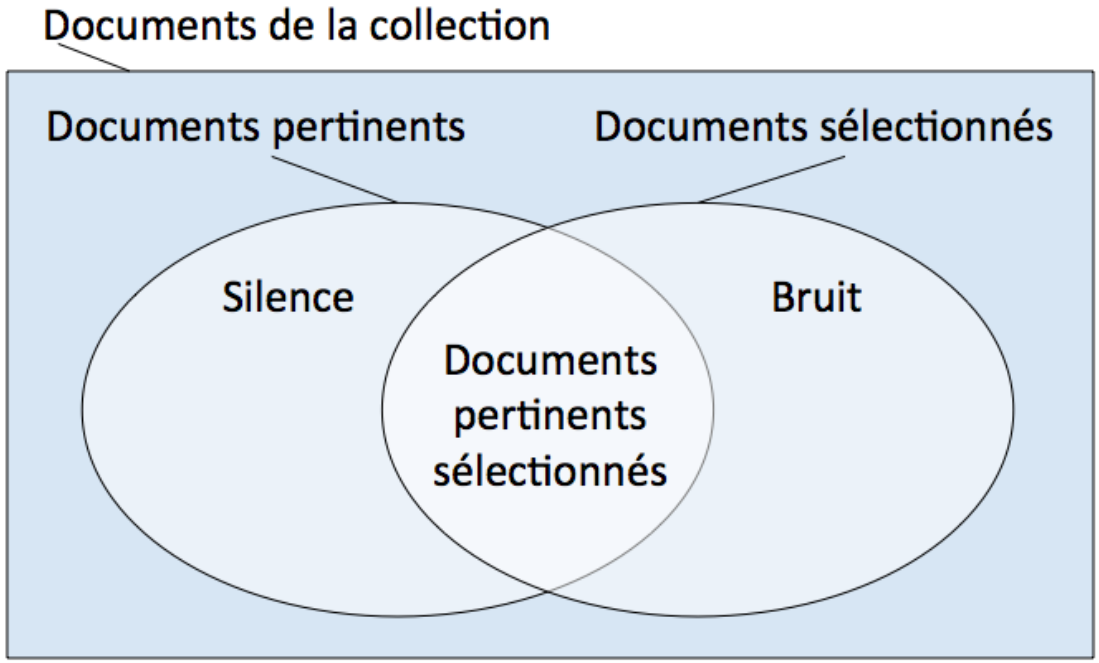


Comment mesurer l'efficacité d'un système ?

- Différents types de mesure
  - Rappel/Précision
  - Orientées rang
  - Prise en compte des degrés de pertinence

Métriques orientées rappel/précision

Mesures orientées rappel/précision



- Les deux mesures les plus courantes sont :
  - Rappel : est-ce que le système retourne TOUS les documents pertinents ?
  - Précision : est-ce que le système retourne QUE les documents pertinents ?

Mesures orientées rappel/précision

	Relevant	Non Relevant
Retrieved	True Positive (tp)	False Positive (fp)
not retrieved	False Negative (fn)	True Negative (tn)

- Précision = capacité à ne retrouver QUE des documents pertinents

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

- Rappel = capacité à retrouver TOUS les documents pertinents

$$Rappel = \frac{tp}{tp + fn} \tag{2}$$

- Remarque : accuracy n'est pas une mesure pour la RI

$$Accuracy = \frac{tp + fn}{tp + fp + tn + fn} \tag{3}$$

Mesures orientées rappel/précision

- Exemple

Rang	Doc	Pertinence
1	324	X
2	654	X
3	454	
4	472	
5	789	X
6	148	X
7	65	
8	32	X
9	78	
10	439	

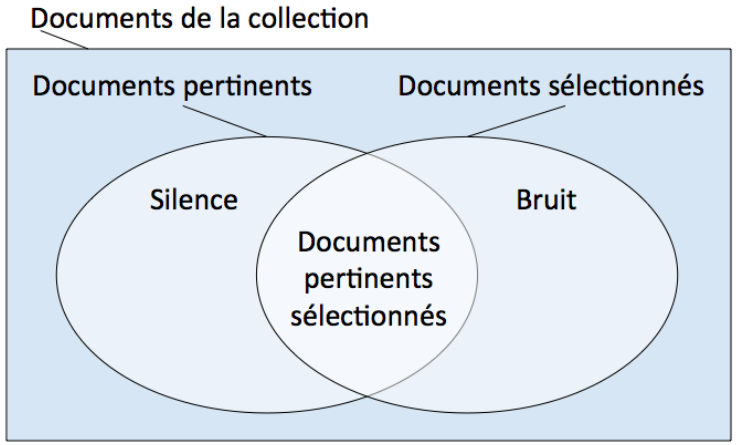
Exercice

Calculer les mesures de rappel et de précision sachant que la collection inclue 10 documents pertinents

Evaluation d'un système

- Calcul des mesures pour chaque requête
- Moyenne sur l'ensemble des requêtes

Mesures orientées rappel/précision



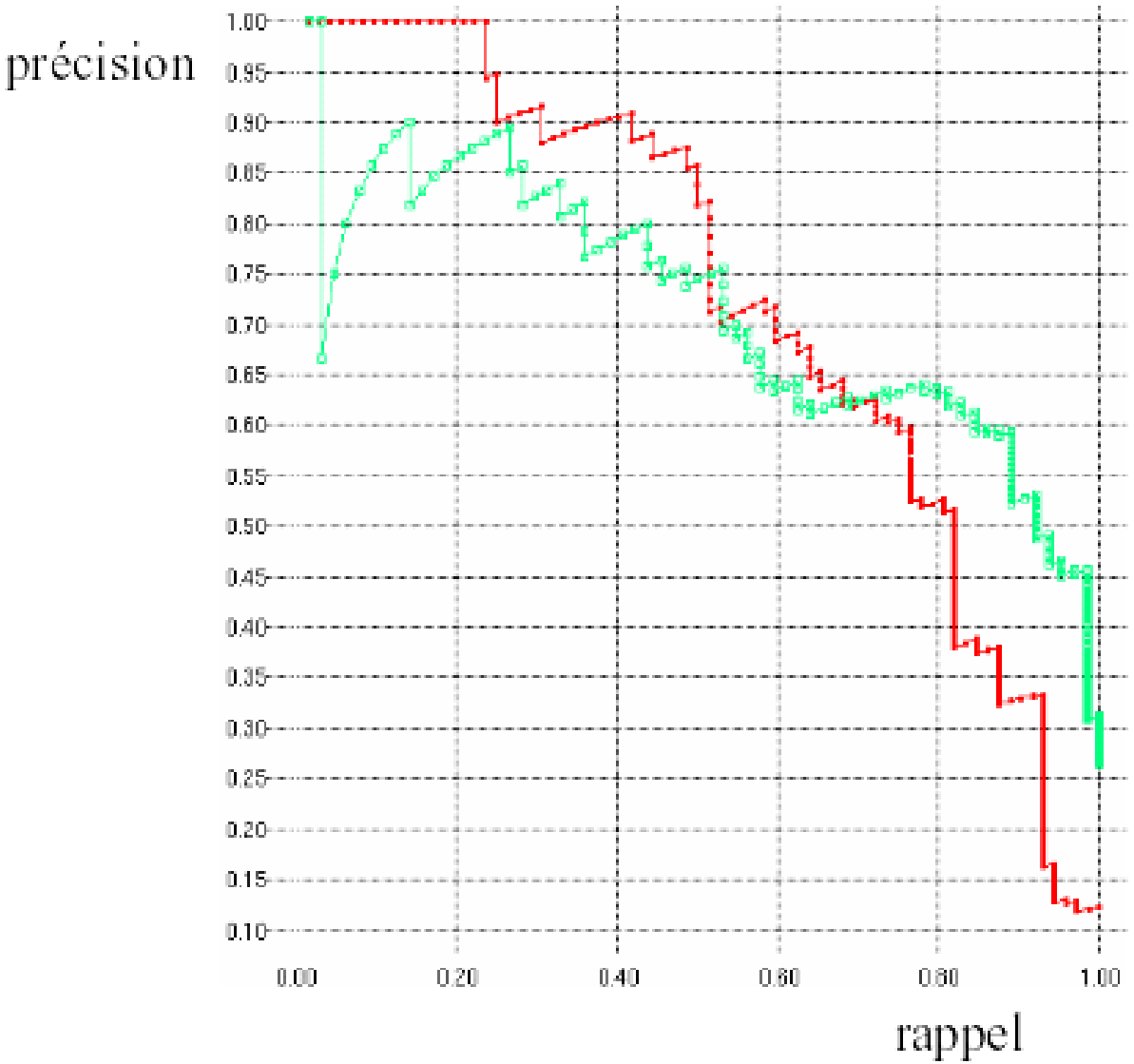
- Rappel et précision sont en général antogonistes
  - Sélection de toute la collection → R=1, P=0
  - Sélection d'un seul document pertinent → R=0, P=1
- Suivant les utilisations, on peut vouloir favoriser précision (e.g. web) ou rappel (documentalistes)
- Mesure qui combine les deux métriques : F-mesure :

$$F_\beta = (1 + \beta^2) \frac{P * R}{\beta^2 P + R} \tag{4}$$

Mesures orientées rappel/précision "Top list"

- Les moteurs de recherche renvoient en général des listes ordonnées : l'idéal est de retourner les documents en tête de liste
- Métriques adaptées
  - Précision à k : $P@k(q) = \frac{1}{k} \sum_{i=1}^{k} R_{d_i,q}$ avec $R_{d_i,q} \in \{0,1\}$ jugement de pertinence pour le document de rang i renvoyé par le système.
  - Rappel à k : $R@k(q) = \frac{1}{|R|} \sum_{i=1}^{k} R_{d_i,q}$ avec $R_{d_i,q} \in \{0,1\}$ jugement de pertinence pour le document de rang i renvoyé par le système.

Courbes de rappel/précision



Courbes de rappel/précision

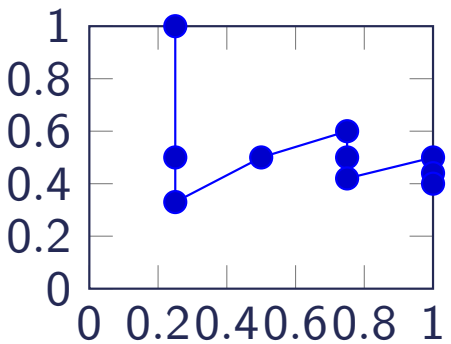
Rg	RSV(q,d)	Pertinence	Rapel	Precision
1	0.95	1		
2	0.82	0		
3	0.75	0		
4	0.7	1		
5	0.65	1		
6	0.5	0		
7	0.4	0		
8	0.35	1		
9	0.2	0		
10	0.1	0		

Exercice

Tracer la courbe de rappel pour cet ordonnancement.

Courbes de rappel/précision

Rg	RSV(q,d)	Pertinence	Rappel	Precision
1	0.95	1	1/4	1
2	0.82	0	1/4	1/2
3	0.75	0	1/4	1/3
4	0.7	1	1/2	1/2
5	0.65	1	3/4	3/5
6	0.5	0	3/4	1/2
7	0.4	0	3/4	3/7
8	0.35	1	1	1/2
9	0.2	0	1	4/9
10	0.1	0	1	2/5



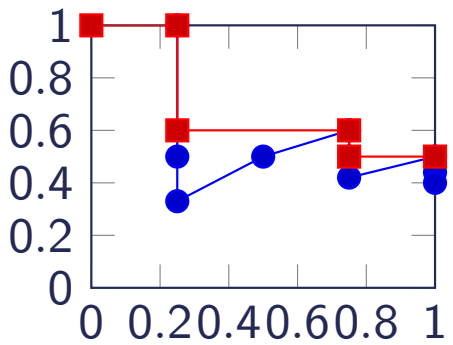
Précision interpolée

La précision interpolée au point de rappel r_j est égale à la valeur maximale des précisions obtenues aux points de rappel r, tel que $r >= r_j$

$$P_{interp}(r) = max_{r' > r} P(r') \quad (5)$$

Courbes de rappel/précision

Rg	RSV(q,d)	Pertinence	Rapel	Precision	Pinterp
1	0.95	1	1/4	1	1
2	0.82	0	1/4	1/2	1
3	0.75	0	1/4	1/3	1
4	0.7	1	1/2	1/2	0.6
5	0.65	1	3/4	3/5	0.6
6	0.5	0	3/4	1/2	0.6
7	0.4	0	3/4	3/7	0.6
8	0.35	1	1	1/2	0.5
9	0.2	0	1	4/9	0.5
10	0.1	0	1	2/5	0.5



Précision interpolée (courbe rouge)

La précision interpolée au point de rappel r_j est égale à la valeur maximale des précisions obtenues aux points de rappel r, tel que $r >= r_j$

$$P_{interp}(r) = max_{r' > r} P(r') \quad (6)$$

avec $r = \{0, 0.1, ..., 1\}$ ou r issu de la distribution de l'ordonnancement.
$\rightarrow$ Estime le pourcentage de documents pertinents qu'un utilisateur observera s'il veut atteindre un rappel au moins égal à r

Précision moyenne

- Précision moyenne (AvgP) est la moyenne des valeurs de précision des documents pertinents par rapport à la requête :

$$AvgP(q) = \frac{1}{n_+^q} \sum_{k=1}^{N} R_{d_k,q} \; x \; P@k(q) \tag{7}$$

  - On peut également calculer la moyenne arithmétique de la précision interposée prise sur 11 points de rappel (approximation de l'aire sous la courbe précision-rappel)
- Moyenne des précisions moyennes (MAP) est la moyenne des AvgP sur l'ensemble des requêtes :

$$MAP = \frac{1}{|Q|} AvgP(q) \tag{8}$$

Métriques orientées rang

Mesures orientées rang

- Hypothèse : les documents pertinents doivent être retournés en premier dans la liste
- Moyenne des rangs inverses (Mean reciprocal rank) : moyenne du rang du premier document sur l'ensemble des requêtes

$$MRR = \frac{1}{|Q|} \sum_{q_h \in Q} \frac{1}{Rank_h} \tag{9}$$

Gain cumulé normalisé

- Discounted cumulative gain (DCG)
  - Utilisé dans le cadre de la recherche Web
  - Utilise une information de pertinence graduée (5 niveaux)
  - Mesure le gain d'information apporté par un document en fonction de sa position dans la liste des résultats
  - Pour la RI Web seules les premières informations présentées sont importantes
- Hypothèses
  - Les documents pertinents sont plus utiles quand ils apparaissent à un rang élevé.
  - Les documents très pertinents sont plus utiles que les peu pertinents qui sont plus utiles que les non pertinents.

Gain cumulé normalisé

- Cumulative Gain (CG) - ancêtre de DCG
  - CG au rang p

$$CG_p = \sum_{i=1}^{p} rel_i \tag{10}$$

  - où rel_i est la pertinence graduée du document i
  - Ne tient pas compte de l'ordre des documents
- Discounted Cumulatige Gain (DCG)
  - Prise en compte de l'ordre des documents par une fonction décroissante du rang

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2(i)} \tag{11}$$

  - Autres formulations possibles

Normalized DCG

- Pour moyenner DCG sur un ensemble de requêtes, on calcule une version normalisée NDCG
  - On suppose que l'on dispose d'une liste idéale de résultats dont le DCG_p vaut $IDCG_p$

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{12}$$

  - On moyenne ensuite sur l'ensemble des requêtes
  - Il faut bien sûr disposer de la liste idéale... $\rightarrow$ petites mains du web...

NDCG - Exercice

Rang	Doc	Pertinence
1	324	1
2	654	2
3	454	
4	472	
5	789	1
6	148	1
7	65	
8	32	2
9	78	
10	439	

Exercice

Calculer la mesure de NDCG pour l'ordonnancement précédent.

Comparaison des systèmes de RI

TopicID	X	Y
01	0.70	0.50
02	0.30	0.10
03	0.20	0.00
04	0.60	0.20
05	0.40	0.40
06	0.40	0.30
07	0.00	0.00
08	0.70	0.50
09	0.10	0.30
10	0.30	0.30
11	0.50	0.40
12	0.40	0.40
13	0.00	0.10
14	0.60	0.40
15	0.50	0.20
16	0.30	0.10
17	0.10	0.10
18	0.50	0.60
19	0.20	0.30
20	0.10	0.20

So you used a test collection that has n=20 topics to compute nDCG scores for two systems X and Y.

Which system is more effective?

Scores for X, Y: $(x_1, \ldots, x_n) \quad (y_1, \ldots, y_n)$

Per-topic difference: $d_j = x_j - y_j$

Sample mean of the differences: $\bar{d} = \sum_{j=1}^{n} d_j / n$

0.0750

Sample variance: $V = \sum_{j=1}^{n} (d_j - \bar{d})^2 / (n-1)$

0.0251

Figure 1 – From Tetsuya Sakai, 2005

Comparaison des systèmes de RI

TopicID	X	Y
01	0.70	0.50
02	0.30	0.10
03	0.20	0.00
04	0.60	0.20
05	0.40	0.40
06	0.40	0.30
07	0.00	0.00
08	0.70	0.50
09	0.10	0.30
10	0.30	0.30
11	0.50	0.40
12	0.40	0.40
13	0.00	0.10
14	0.60	0.40
15	0.50	0.20
16	0.30	0.10
17	0.10	0.10
18	0.50	0.60
19	0.20	0.30
20	0.10	0.20

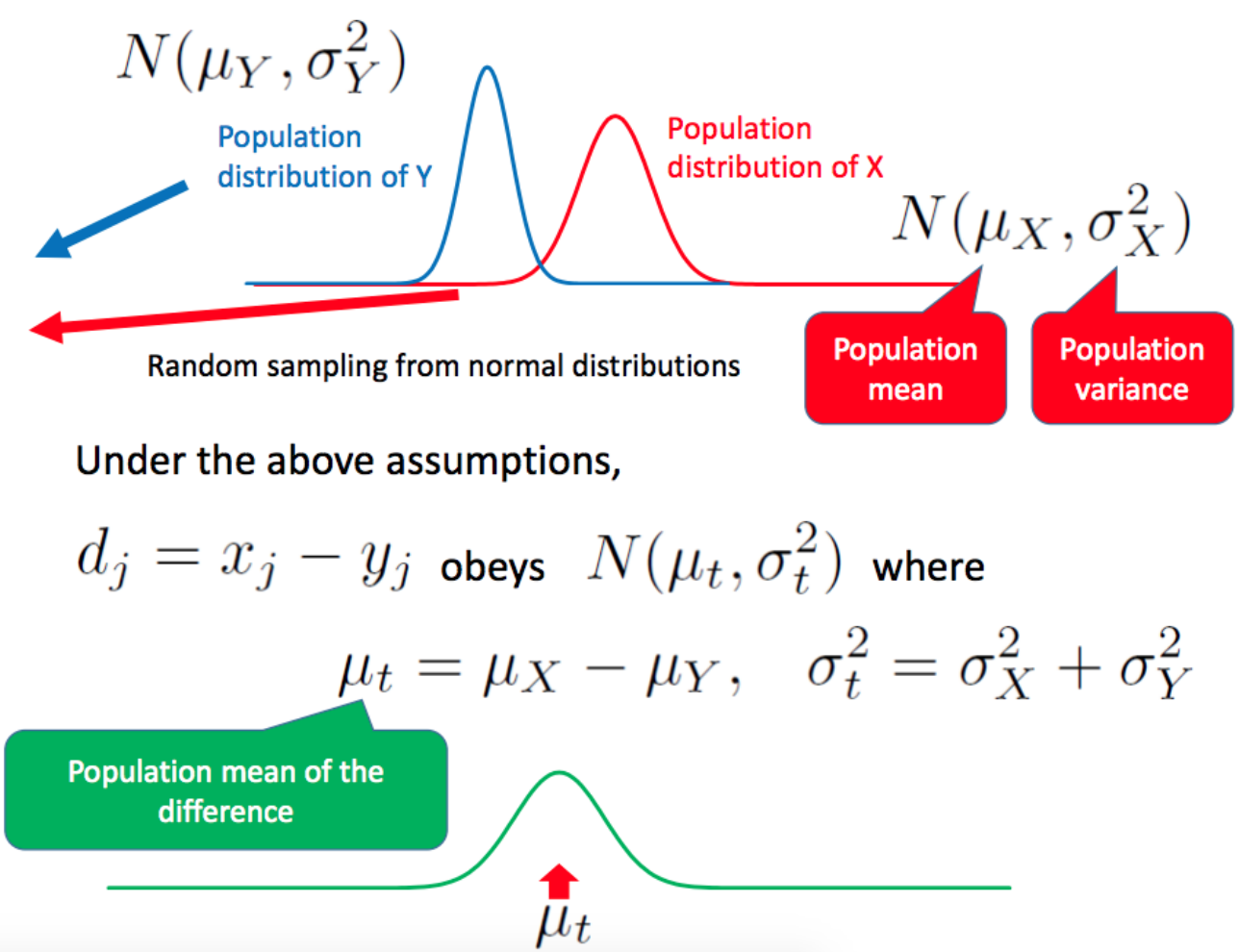


Figure 2 – From Tetsuya Sakai, 2005

Comparaison des systèmes de RI

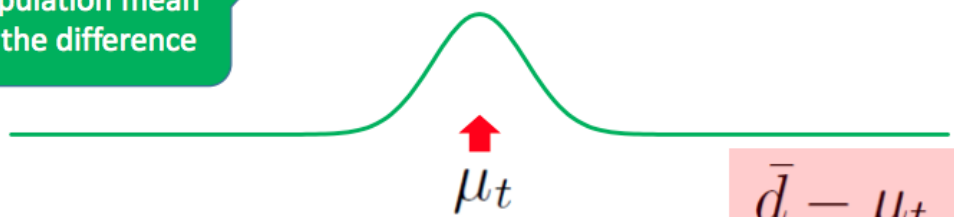
TopicID	X	Y
01	0.70	0.50
02	0.30	0.10
03	0.20	0.00
04	0.60	0.20
05	0.40	0.40
06	0.40	0.30
07	0.00	0.00
08	0.70	0.50
09	0.10	0.30
10	0.30	0.30
11	0.50	0.40
12	0.40	0.40
13	0.00	0.10
14	0.60	0.40
15	0.50	0.20
16	0.30	0.10
17	0.10	0.10
18	0.50	0.60
19	0.20	0.30
20	0.10	0.20

Under the above assumptions,

$d_j = x_j - y_j$ obeys $N(\mu_t, \sigma_t^2)$ where

$\mu_t = \mu_X - \mu_Y, \quad \sigma_t^2 = \sigma_X^2 + \sigma_Y^2$

Population mean of the difference



$$\frac{\bar{d} - \mu_t}{\sqrt{\sigma_t^2/n}}$$ obeys $N(0, 1^2)$

Which system is more effective?

Or, which of these hypotheses is true?

$H_0 : \mu_t = 0, \quad H_1 : \mu_t \neq 0$

If you look at the populations, X and Y are equally effective

If you look at the populations, X and Y are actually different

Figure 3 – From Tetsuya Sakai, 2005

Comparaison des systèmes de RI

TopicID	X	Y
01	0.70	0.50
02	0.30	0.10
03	0.20	0.00
04	0.60	0.20
05	0.40	0.40
06	0.40	0.30
07	0.00	0.00
08	0.70	0.50
09	0.10	0.30
10	0.30	0.30
11	0.50	0.40
12	0.40	0.40
13	0.00	0.10
14	0.60	0.40
15	0.50	0.20
16	0.30	0.10
17	0.10	0.10
18	0.50	0.60
19	0.20	0.30
20	0.10	0.20

$$H_0 : \mu_t = 0, \quad H_1 : \mu_t \neq 0$$

If Ho is true, the t statistic obeys a t distribution with ϕ=(n-1) degrees of freedom.

Significance level α: areas under curve = a pre-determined probability (e.g. 5%) of observing something very rare

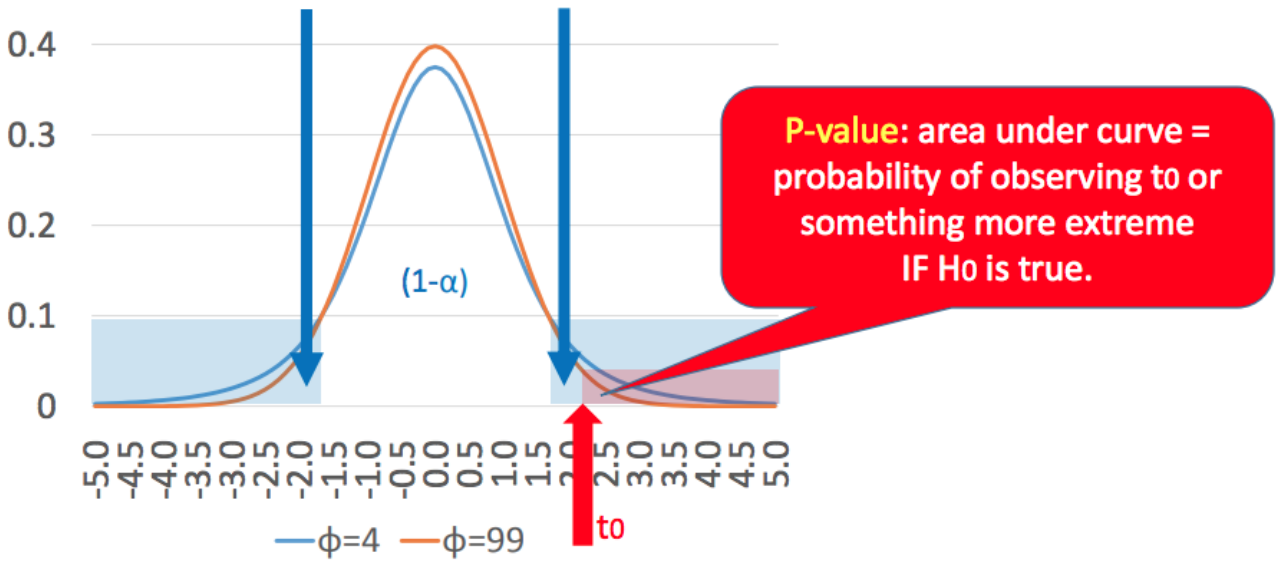


Figure 4 – From Tetsuya Sakai, 2005

Comparaison des systèmes de RI

- Tests de significativité
  - Vérifier si la différence de performances entre deux systèmes est significative

    Rejeter H0 (pas de différence entre A et B)
    Accepter H1 (A et B sont différents)
  - Plusieurs tests : Student, Wilcoxon, ...
  - Dépend d'un paramètre α

    $0.05 \rightarrow 0.95\%$ de chance que les systèmes soient différents

Comparaison des systèmes de RI

- Exemple de présentation de résultats

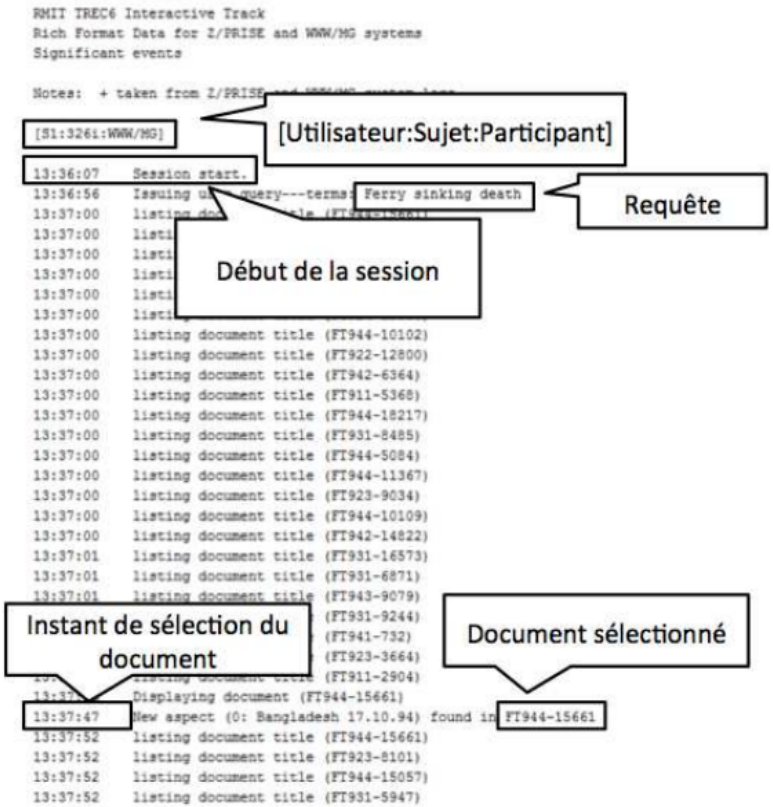
Apprentissage → *Evaluation*		$Prec^S$@20 value	$Prec^S$@20 $\%Tx$	$Rappel^S$@20 value	$Rappel^S$@20 $\%Tx$	F^S@20 value	F^S@20 $\%Tx$
US2	BM25-RIC	0,016	+185,64***	0,019	+139,62***	0,0177	+166.71***
	Logit-RIC	0,038	+21,66	0,031	+43.50 *	0,033	+31.75 *
→	GS-RIC	0,015	+204,44 ***	0,008	+429.17 ***	0,009	+345.81 ***
SansRole	PM-RIC	0,019	+136,62 ***	0,006	+719.35 ***	0,008	+432.37 ***
	MineRank(q)	**0.046**		**0.045**		**0.044**	
	MineRank(t)	0.040		0.040		0.040	
SansRole	BM25-RIC	0,075	-5,00	0,063	+335,58	0,069	+63,27
	Logit-CIT	0,071	+0,33	0,266	+3,76 ***	0,111	+1,33 ***
→	GS-RIC	0,058	+23,76	0,039	+595,56 *	0,046	+142,58
US2	PM-RIC	0,092	-22,83	0,078	+254,57	0,084	+32,86
	MineRank(q)	**0.071**		**0.276**		**0.112**	
	MineRank(t)	0.064		0.238		0.112	

Autres protocoles d'évaluation...

- Quand l'humain rentre en jeu...
  - Evaluation basée sur les logs utilisateurs
  - User-study

Autres protocoles d'évaluation

- Evaluation basée sur les logs utilisateurs



Autres protocoles d'évaluation

- Evaluation basée sur les logs utilisateurs
  - Permet d'appliquer a posteriori des modèles sur des données utilisateurs

  - Avantages

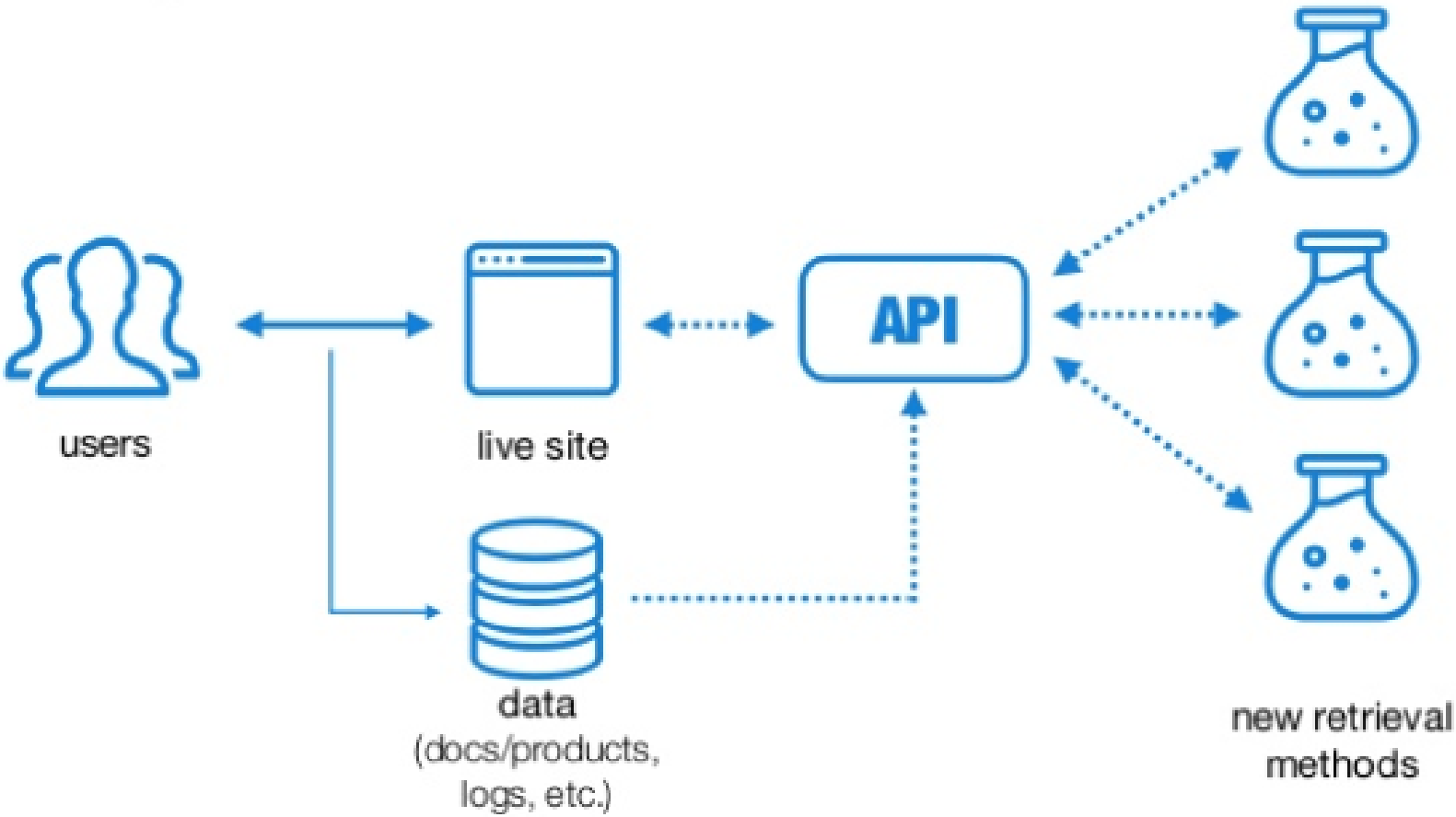
    Besoin en information généré par un utilisateur réel
    Automatisation de l'évaluation

  - Inconvénients

    "Artificiel"

Autres protocoles d'évaluation

- User-study / Living labs



K. Balog, L. Kelly, and A. Schuth. **Head First: Living Labs for Ad-hoc Search Evaluation.** *CIKM'14*

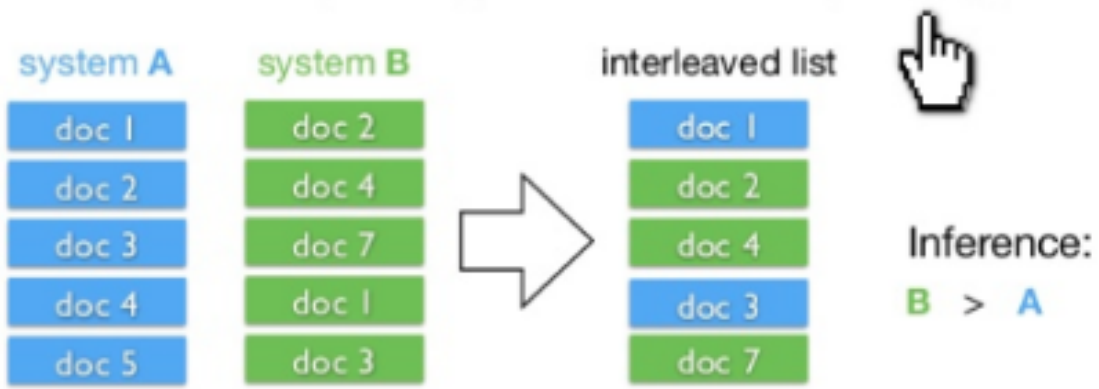
Autres protocoles d'évaluation

- User-study
  - Utilisateur interagit en temps réel avec le système
- Avantages
  - Evaluation directe du système
  - Au plus proche de l'utilisateur !
- Inconvénients
  - Collecte fastidieuse, coûteuse, ...
  - Difficile d'évaluer toutes les variantes d'un modèle (paramétrage, etc. . .)
  - Evaluation de plusieurs modèles ?

Evaluation des expérimentations en temps réel : Interleaving et A/B tests

Interleaving

- Site provides the **set of candidate items** that can be re-ranked (safety mechanism)
- Experimental ranking is **interleaved** with the production ranking
  - Meeds 1-2 order of magnitudes data than A/B testing (also, it is within subject as opposed to between subject design)



A vous de jouer... TD is coming !