Chapitre 5 :
Evaluation des performances dans les SRIs

1

1. Objectif

L'objectif de cette évaluation en RI est la comparaison entre des SRI . On ne mesure pas la performance absolue d'un SRI car non significative, mais on mesure la performance relative d'un SRI par rapport à un autre.

2. Démarche d'évaluation

2.1. Démarche analytique (formelle)

Cette démarche se base sur des preuves mathématiques afin de déterminer la meilleure approche de RI ou le meilleur SRI. Mais, en RI, il y a plusieurs facteurs (l'indexation, la pertinence, la requête, la distribution des termes, …etc.), qui sont difficiles à formaliser mathématiquement. Donc, cette démarche ne peut pas être utilisée en RI.

2

2.2. Démarche expérimentale (benchmarking)

Cette démarche est basée sur un environnement de tests et des mesures d'évaluation. Cette démarche est très utilisée en RI.

2.2.1. Environnement de test

Un environnement de test ou d'évaluation doit contenir au minimum les informations suivantes :

- Un ensemble de documents (collection de documents)
- Un ensemble de requêtes de tests
- Les documents pertinents pour chaque requête de test.

Il existe plusieurs environnement de test pour la RI, à savoir :
CACM, CISI, CRAN, MED, TIME, TREC

3

3. Les mesures d'évaluation

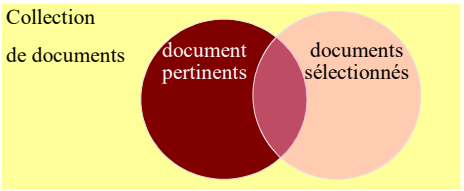
Il existe plusieurs mesure d'évaluation des SRI.

3.1. Le Rappel et la précision

- **Rappel :** La capacité d'un système à sélectionner **tous** les documents pertinents de la collection.
- **Précision :** La capacité d'un système à sélectionner **que** des documents pertinents

4

Le Rappel et la précision



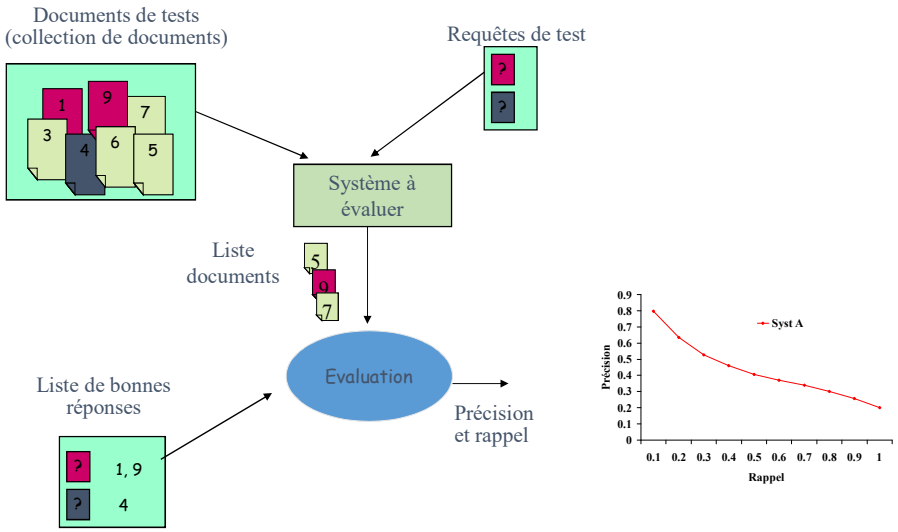
	Sélectionnés	Non sélectionnés
Non pertinents	Sélection. & Non Pert.	Non sélection. & Non Pert.
Pertinents	Sélection. & Pert	not sélection. mais Pert.

$$rappel = \frac{Nombre\ de\ documents\ pertinents\ séléctionnés}{Nombre\ total\ de\ documents\ pertinents}$$

$$précision = \frac{Nombre\ de\ documents\ pertinents\ \ sélectionnés}{Nombre\ total\ de\ documents\ sélectionnés}$$

5

Processus de calcul du Rappel et de la précision



6

Exercice :

Soit deux systèmes de recherche d'information A et B évalués sur une liste de 10 documents {d1, d2, d3, d4, d5, d6, d7, d8, d9, d10}. On sait que les documents d1, d4, d6 et d10 sont pertinents et les autres ne le sont pas (selon l'environnement de tests).

- Le système A retourne les documents d5, d1, d6, d2
- Le système B retourne les documents d7, d8, d1, d6, d2, d10, d9

Calculer la précision et le rappel pour les deux systèmes A et B. Le quel des deux systèmes est meilleur ?

Solution :

$Rappel_A = 2/4 = 0.5$
$Précision_A = 2/4 = 0.5$

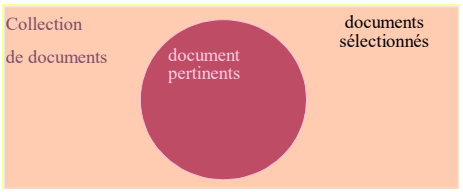
$Rappel_B = 3/4 = 0.75$
$Précision_B = 3/7 = 0.428$

On remarque que :

- du point de vu rappel le système B qui est meilleur
- du point de vu précision c'est le système A qui meilleur

7

Pourquoi deux facteurs ?



- FACILE de faire du rappel il suffit de sélectionner toute la collection
- MAIS, la précision sera très faible

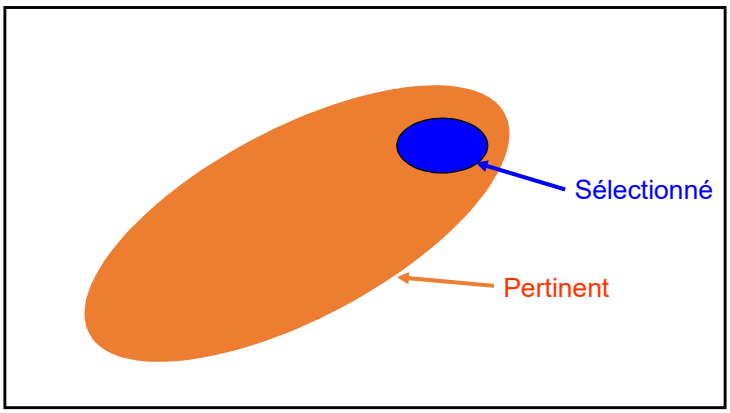
8

Pertinent vs. Sélectionné

9

Sélectionné vs. Pertinent

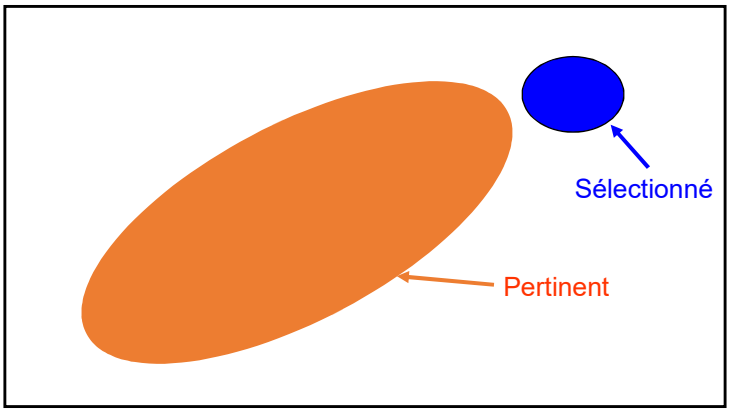
Précision très élevée, rappel très faible



10

Sélectionné vs. Pertinent

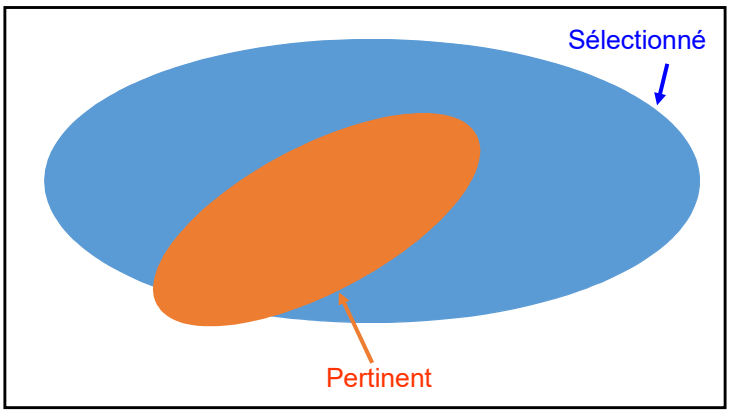
Précision très faible, rappel très faible (en fait, 0)



11

Sélectionné vs. Pertinent

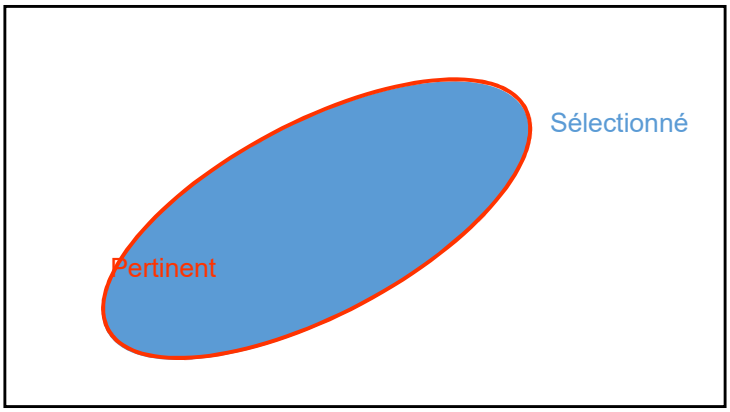
Rappel élevé, mais précision faible



12

Sélectionné vs. Pertinent

Précision élevée, rappel élevé (idéal, mais difficile)



13

3.1.1. Calcul du rappel et la précision à chaque document pertinent du système de RI

Pour chaque requêtes de test :

1. Lancer chaque requête sur la collection de tests.
2. Marquer ses documents pertinents par rapport à la liste de test.
3. Calculer le rappel et la précision pour chaque document pertinent de sa liste des résultats

14

Exemple : (pour une requête)

Le nombre total de documents pertinents est = 6

Résultat du système à évaluer

n	doc #	Pertinent
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=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

Il manque un document pertinent. On atteindra pas le 100% de rappel

15

R	Pr
0.167	1
0.333	1
0.5	0.75
0.667	0.667
0.833	0.38

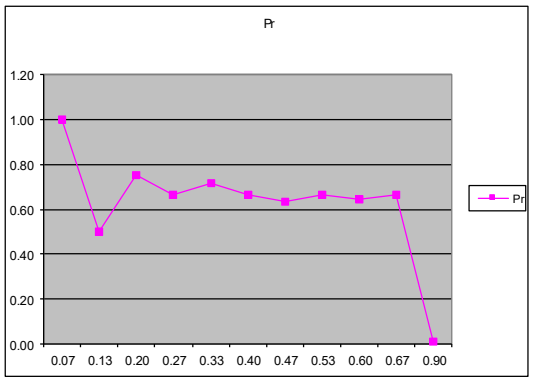
R	Pr
0.0	1
0.1	1
0.2	1
0.3	1
0.4	0.75
0.5	0.75
0.6	0.667
0.7	0.38
0.8	0.38
0.9	0
1	0

16

Courbe rappel/précision pour une requête à chaque document pertinent

Exemple : (pour une requête)

Ra	Pr
0,07	1,00
0,13	0,50
0,20	0,75
0,27	0,67
0,33	0,71
0,40	0,67
0,47	0,64
0,53	0,67
0,60	0,64
0,67	0,67
0,90	0,01



17

Interpolation de la courbe Rappel/Précision

Interpoler une précision pour chaque point de rappel :

$r_j \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$

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 $r >= r_j$

$$P(r_j) = \max_{r \geq r_j} P(r)$$

18

Exemple : Interpolation des Précisions

$$P(r_j) = \max_{r \geq r_j} P(r)$$

Ra	Pr
0,07	1,00
0,13	0,50
0,20	0,75
0,27	0,67
0,33	0,71
0,40	0,67
0,47	0,64
0,53	0,67
0,60	0,64
0,67	0,67
0,90	0,01

Interpolation de la précision à chaque points du Rappel →

Ra	Pr
0,0	1
0,1	0,75
0,2	0,75
0,3	0,71
0,4	0,67
0,5	0,67
0,6	0,67
0,7	0,01
0,8	0,01
0,9	0,01
1	0

19

3.2. Précision moyenne pour une requête

On souhaite souvent avoir une valeur unique, par exemple pour les algorithmes d'apprentissage pour contrôler l'amélioration.

3.2.1. Précision moyenne non interpolée (PrecAvg) :

Calculer la précisions à chaque apparition d'un document pertinent, puis diviser leur somme sur le nombre de documents pertinents donnés par l'environnement de tests.

20

Exemple : (Précision moyenne non interpolée « AvgPrec » pour une requête)

Le nombre total de document pertinent donnés par l'environnement de test est = 8

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=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

AvgPrec=(1+1+0,75+0,667+0,38)/8

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

21

Exemple de résultats renvoyés par le Programme TREC_EVAL

Total number of documents over all queries
Retrieved: 1000
Relevant: 80
Rel_ret: 30
Interpolated Recall - Precision Averages:
at 0.00 0.4587
at 0.10 0.3275
at 0.20 0.2381
at 0.30 0.1828
at 0.40 0.1342
at 0.50 0.1197
at 0.60 0.0635
at 0.70 0.0493
at 0.80 0.0350
at 0.90 0.0221
at 1.00 0.0150
Average precision (non-interpolated) for all rel docs:
0.1311

22

3.2.2. Autres mesures de moyennes

F-Mesure

- Mesure tenant compte à la fois du rappel et de la précision.
- Introduite par van Rijbergen, 1979
- Moyenne harmonique entre R et P

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}}$$

E-Mesure (F-Mesure paramétrique)

- Une variante de F-Mesure qui tient compte du poids accordé à la précision vis-à-vis du rappel

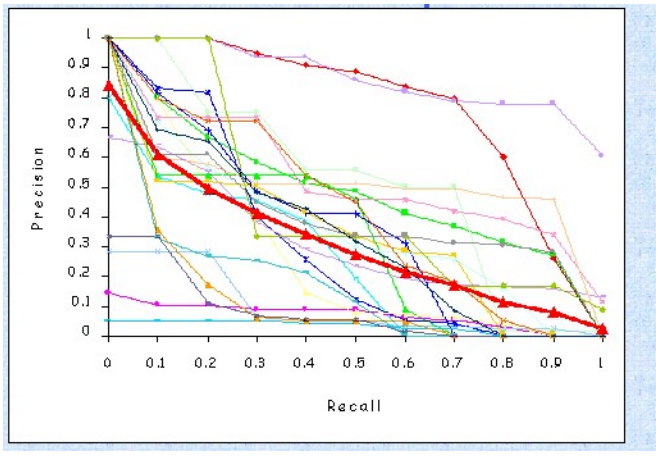
$$E = \frac{(1+\beta^2)PR}{\beta^2 P+R} = \frac{(1+\beta^2)}{\frac{\beta^2}{R}+\frac{1}{P}}$$

β contrôle le compromis R, P:

- β = 1: même poids précision et Rappel (E=F).
- β > 1: Ajout de poids à la précision
- β < 1: Ajout de poids au rappel.

23

3.3. R-P courbes sur l'ensemble des requêtes



Illisible, difficile de comparer deux approches/systèmes requête par requête. On a besoin d'une moyenne entre les requêtes

24

3.4. Moyenne sur plusieurs requêtes

Calculer la moyenne sur plusieurs requêtes :

- **Micro-moyenne (non interpolée) :** chaque document pertinent sur l'ensembles des requêtes est un point de la moyenne.
- **Macro-moyenne (interpolée) :** Calculer la précision moyenne à chaque point de rappel (précision interpolée) pour l'ensemble des requêtes.
- **Moyenne des moyennes :** calculer la moyenne des précisions moyennes

25

Exemple 1 : moyenne des précisions non interpolées pour plusieurs requêtes

Requête1

Ra	Pr
0.07	1.00
0.13	0.50
0.20	0.75
0.27	0.67
0.33	0.71
0.40	0.67
0.47	0.64
0.53	0.67
0.60	0.64
0.67	0.67
0.73	0.01

Requête 2

Ra	Pr
0.07	1.00
0.13	0.50
0.20	0.43
0.27	0.44
0.33	0.45
0.40	0.46
0.47	0.47
0.53	0.47
0.60	0.50
0.67	0.48
0.73	0.22

Micro-moyenne (non interpolée)

1.00
0.50
0.59
0.56
0.58
0.57
0.56
0.57
0.57
0.58
0.12

Supposant que la requête1 a 15 docs pertinents (d'après l'environnement de tests), alors : AvgPrec de cette requête = 6.93/15= 0.462

Supposant que la requête2 a 15 docs pertinents (d'après l'environnement de tests), alors : AvgPrec de cette requête = 5.42/15= 0.361

La moyenne des moyennes des précisions non interpolées est : MoyAvecPrec = (0.462+0.361)/2=0.412

26

Exemple 2 : moyenne des précisions interpolées pour plusieurs requêtes

Requête 1

R	Pr
0	0,629
0,1	0,451
0,2	0,393
0,3	0,3243
0,4	0,271
0,5	0,2424
0,6	0,164
0,7	0,134
0,8	0,09
0,9	0,04
1	0,031

Requête 2

R	Pr
0	0,5017
0,1	0,332
0,2	0,248
0,3	0,171
0,4	0,155
0,5	0,125
0,6	0,089
0,7	0,056
0,8	0,032
0,9	0,027
1	0,02

précision moyenne à chaque point de rappel

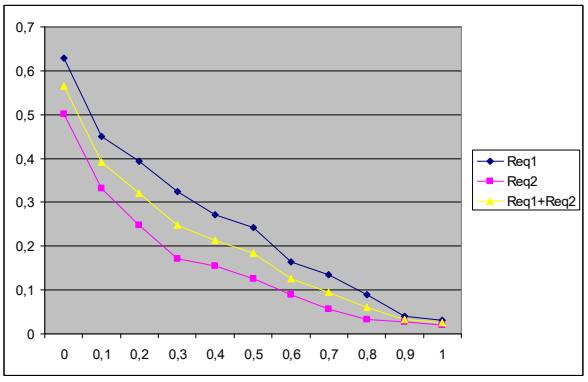
R	Pr
0	0,56535
0,1	0,3915
0,2	0,3205
0,3	0,24765
0,4	0,213
0,5	0,1837
0,6	0,1265
0,7	0,095
0,8	0,061
0,9	0,0335
1	0,0255

Moyenne des moyennes des précision interpolées pour les deux requêtes =(0.56535+0,3915+ 0,3205+ 0,24765+0,213+0,1837 +0,1265+0,095+0,061+0,0335+0,0255)/11

27

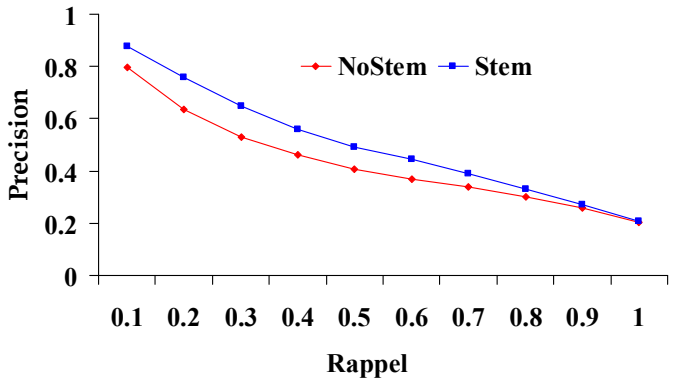
Courbes :

- Rappel/ précisions interpolées requête1
- Rappel/ précisions interpolées requête2
- Rapel/moyenne des précisions interpolées pour plusieurs requêtes



28

Courbes : de deux systèmes du Rapel/moyenne des précisions interpolées pour plusieurs requêtes



29

Mesures focalisées sur le "top" de la liste

30

4. Mesures focalisées sur le "top" de la liste

Dans les cas où :

- Les utilisateurs se focalisent davantage sur les documents pertinents se trouvant en “top” des résultats
- La mesure de rappel n’est pas toujours appropriée : comme dans stratégies de recherche pour lesquelles il y a une réponse unique (navigational search, question answering)

La solution pour ces cas est de mesurer plutôt la capacité d’un SRI à trouver les documents pertinents en top de la liste, parmi ces mesures on a :

- ✓ *Precision au Rang X (Precision at rank X)*
- ✓ *R-Précision (R-Precision)*
- ✓ *Rang réciproque (Reciprocal Rank)*
- ✓ *Gain Cumulé (Discounted Cumulative Gain)*
- ✓ *Gain Cumulé Normalisé (Normalized Discounted Cumulative Gain)*

31

4.1. Precision au Rang X (Precision at rank X)

On calcule la précision à différent niveau de documents, comme : Précision calculée à 5 docs, 10 docs, 15docs, …

Exemple :

1	588	
2	589	
3	576	x
4	590	
5	986	x
6	592	x
7	984	
8	988	
9	578	x
10	985	x
11	103	
12	591	
13	772	x
14	456	
15	990	

Précision à 5 docs = 2/5
Précision à 10 docs = 5/10
Précision à 10 docs = 6/15

32

4.2. R-Précision (R-Precision)

Une façon de calculer une valeur de précision unique : précision au R ème document de la liste des documents sélectionnés par la requête ayant R documents pertinents dans la collection.

Exemple :

1	588	
2	589	
3	576	x
4	590	
5	986	x
6	592	x
7	984	
8	988	
9	578	x
10	985	x
11	103	
12	591	
13	772	x
14	456	
15	990	

Selon l'environnement de tests, il y a 8 documents pertinents, donc R=8, alors :

R-Precision = 3/8 = 0,375

33

4.3. Rang réciproque (Reciprocal Rank)

On calcule l'inverse du rang du premier document pertinent sélectionné

Exemple :

1	588	
2	589	
3	576	x
4	590	
5	986	x
6	592	x
7	984	
8	988	
9	578	x
10	985	x
11	103	
12	591	
13	772	x
14	456	
15	990	

Reciprocal Rank = 1/3 = 0,333

34

4.4. Le Gain Cumulé (Discounted Cumulative Gain)

Le gain cumulé permet de détecter à quel point les documents pertients sont bien ordonés dans la liste des résultats du système à évaluer, selon leurs pertinences de l'environnement de tests.

Le gain cumulé au rang p (DCG_p) est calculé comme suit :

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

Avec :

reli : la similarité donnée par l'environnement de test du document i

i : la position du document dans les résultats du système à évaluer

35

Exemple

Soit une liste de documents retournés par un système A :

d1, d2, d3, d4, d5, d6, d7, d8, d9, d10

Les degrés de pertinence des documents donnés par l'environnement de tests sont :
rsv(d1) =0.3, rsv(d2) =0.2, rsv(d3) =0.3, rsv(d6) =0.4, rsv(d7) =0.5, rsv(d9) =0.3

DCG_{10} = 0.3 + 0.2/1 + 0.3/1.585 + 0 + 0 + 0.4/2.585 + 0.5/2.807 + 0 + 0.3/3.170 + 0
= 1.117

36

4.5. DCG Normalizé (NDCG)

Les valeurs de DCG sont souvent normalisées selon la valeur DGC du classement parfait (DCGIdeal).

NDCG = DCG/DCGIdeal

Exemple :

Pour l'exemple pécédent :

Ordre ideal des documents pertinents selon l'environnement de tests est : d7 d6 d1 d3 d9 d2

$DCG\ Ideal_{10} = 0.5 + 0.4/1 + 0.3/1.585 + 0.3/2.585 + 0.3/2.807 + 0.2/3.170$

$= 1.735$

$NDGC_{10} = 1.117 / 1.735$

$= 0.644$

37

5. La comparaison des systèmes

- Comparer les performances en termes de mesures d'évaluation de deux systèmes A et B :
  - On calcule : (Val(A)-Val(B)/Val(B))*100
  - A partir de 5% on peut considérer que A et meilleur que B
- Comparer leurs courbes
  - La courbe de A est toujours supérieure à celle de B

Remarque :

Cette comparaison est relative juste à une collection. Que se passe t-il quand on change de collection ?

38

6. Avantages et inconvénients des collections de tests

6.1. Avantages

- Mesures de performances
- Possibilité de comparaison avec d'autres travaux

6.2. Inconvénients

- Les résultats obtenus sont propres à la collection.
- Ne répondent pas à toutes les tâches de RI, notamment celles orientées utilisateur