Evaluation des performances dans les SRI

Qu'est ce qui marche ?

LTC
BM25
Statistique vs. linguistique
Indexer les passages
Probabiliste
Connexionnisme
Mots vides
thésaurus
IDF
tf
phrases
LTU
Groupes de mot² s
liens web
Vectoriel

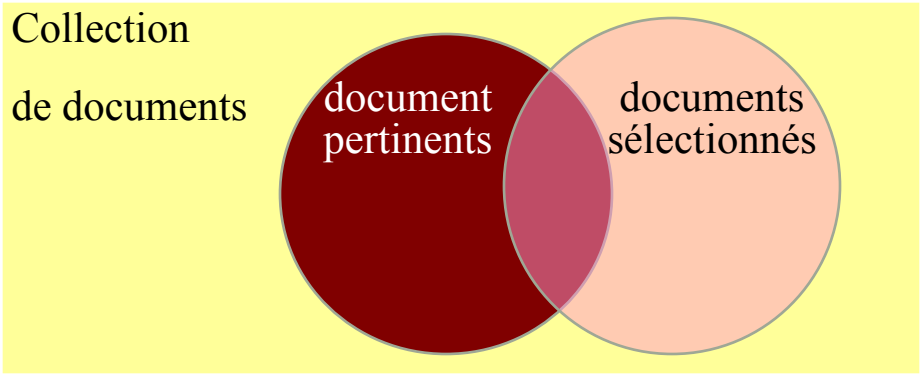
Evaluer

Evaluation

- Identifier les critères (Cleverdon 66)
  - Facilité d'utilisation du système
  - Coût accès/stockage
  - Présentation des résultats
  - Capacité d'un système à sélectionner des documents pertinents.

> Rappel : capacité d'un système à sélectionner tous les documents pertinents de la collection
>
> Précision : capacité d'un système à sélectionner que des documents pertinents

Précision et Rappel

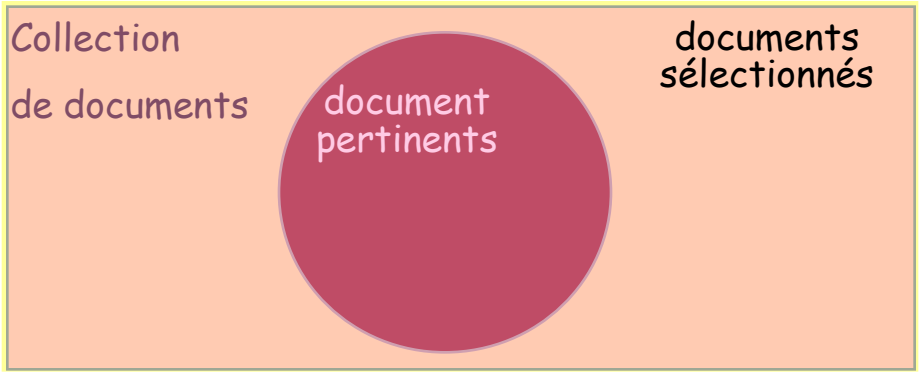


	retrieved	not retrieved
irrelevant	Sélection. & Non Pert.	Non sélection. & Non Pert.
relevant	Sélection. & Pert	not sélection. mais Pert.

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

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

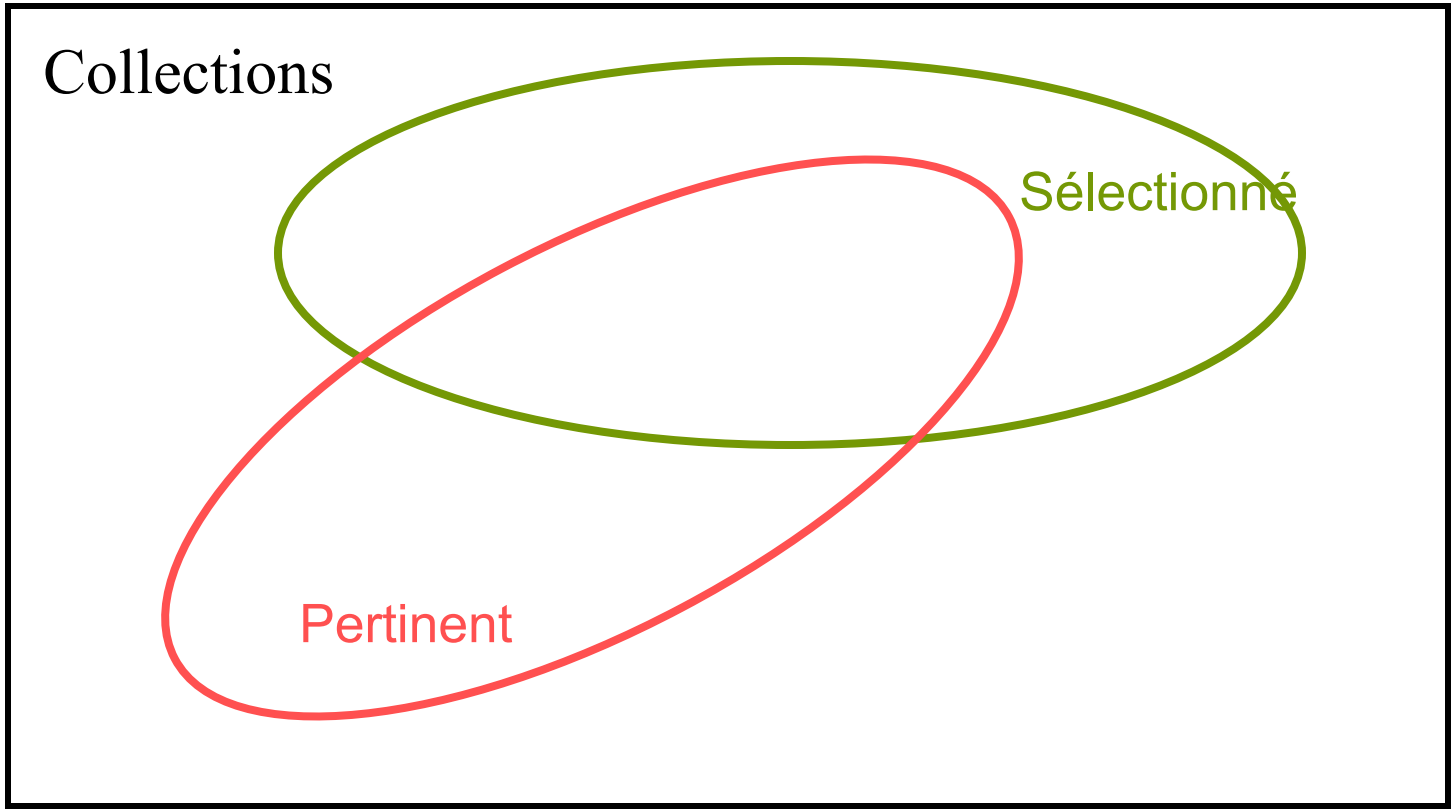
Pourquoi deux facteurs ?



- FACILE de faire du rappel il suffit de sélectionner toute la collection

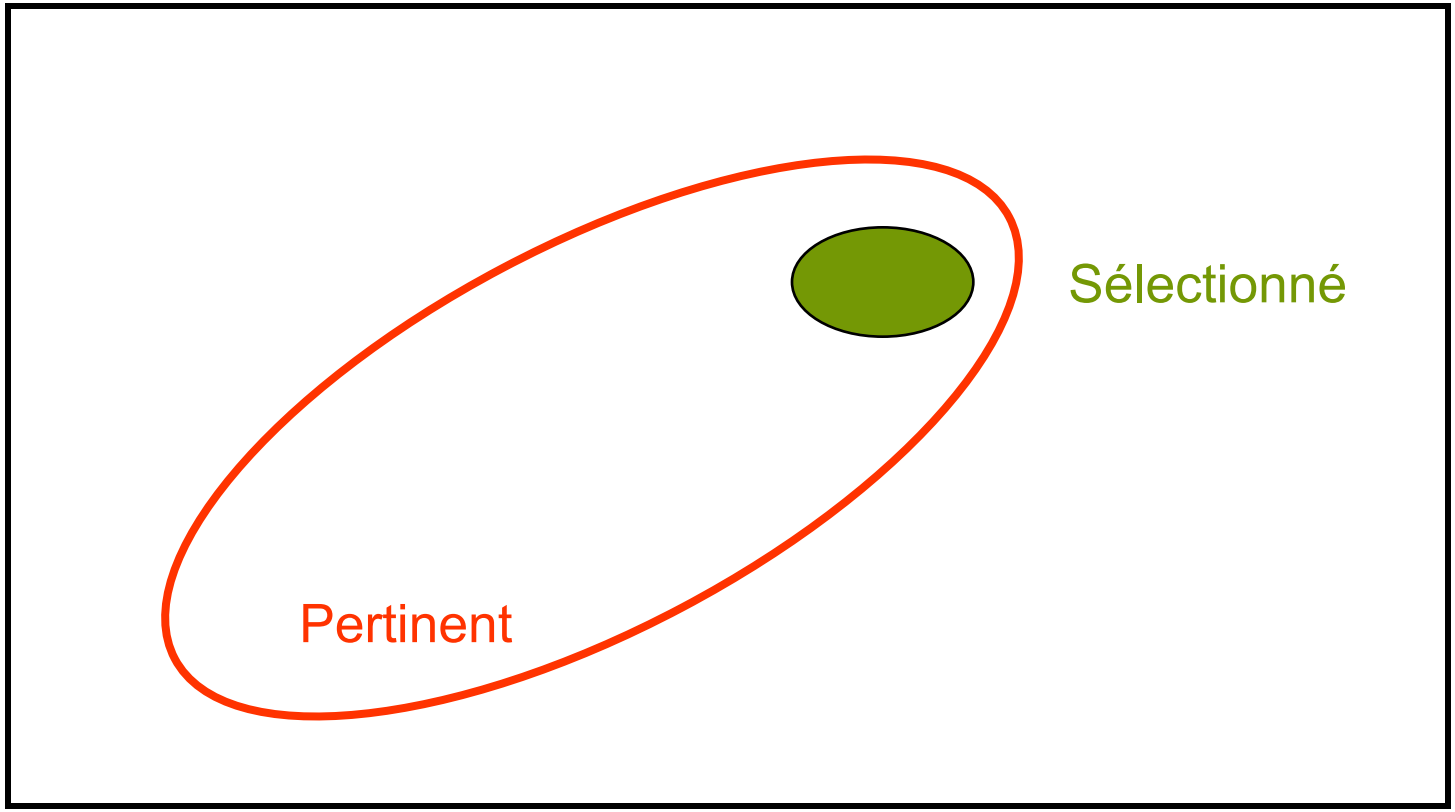
- MAIS, la précision sera très faible

Pertinent vs. Sélectionné



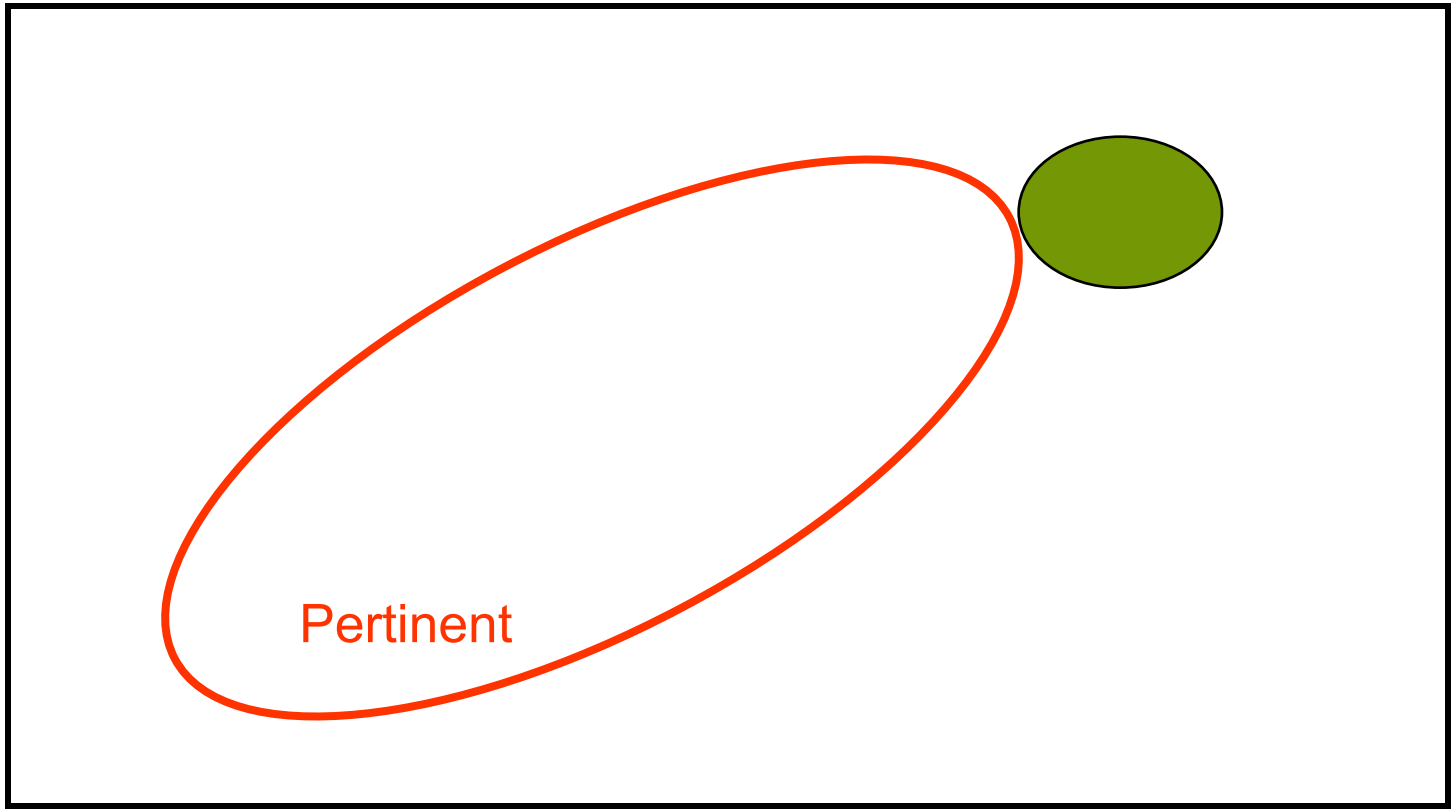
Sélectionné vs. Pertinent

Précision très élevée, rappel très faible



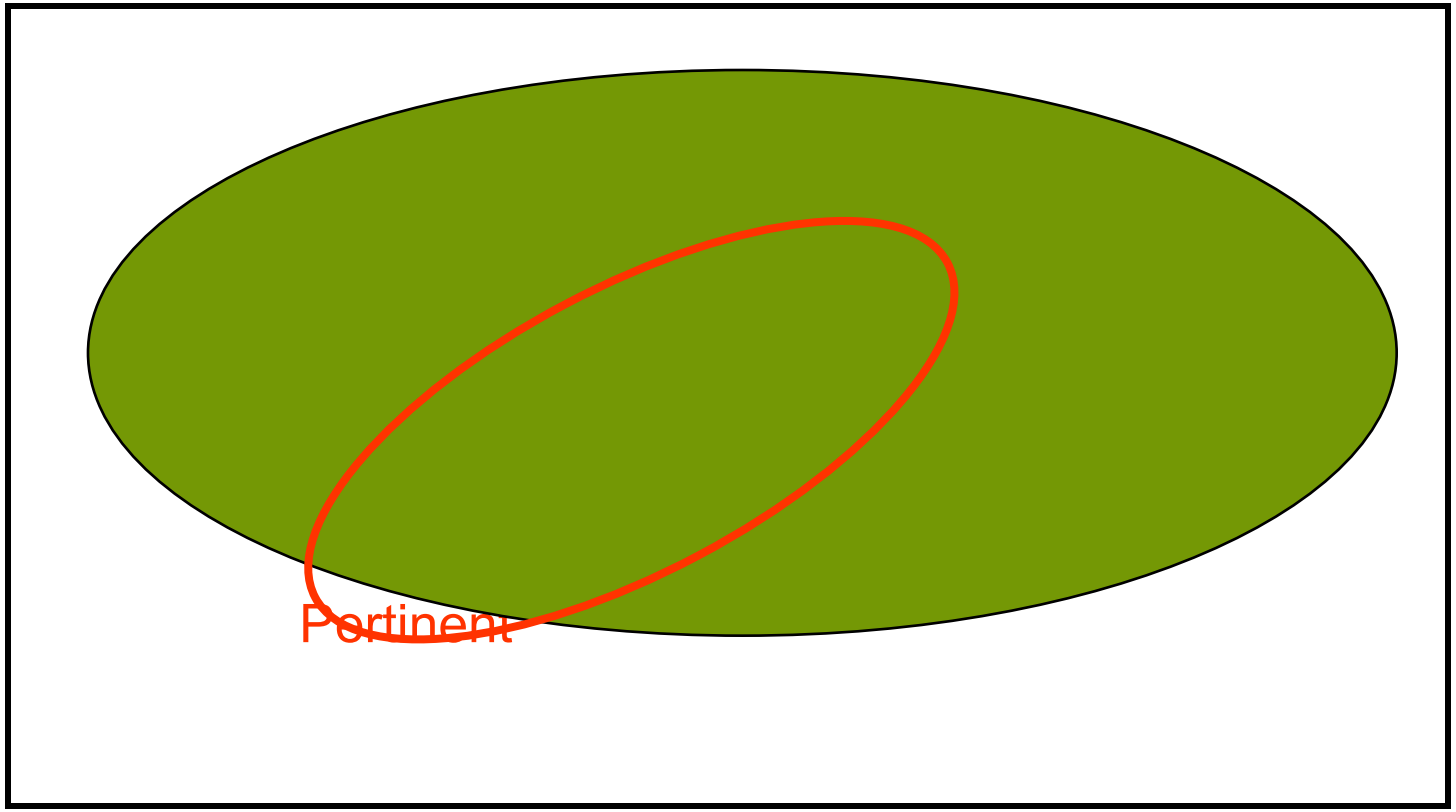
Sélectionné vs. Pertinent

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



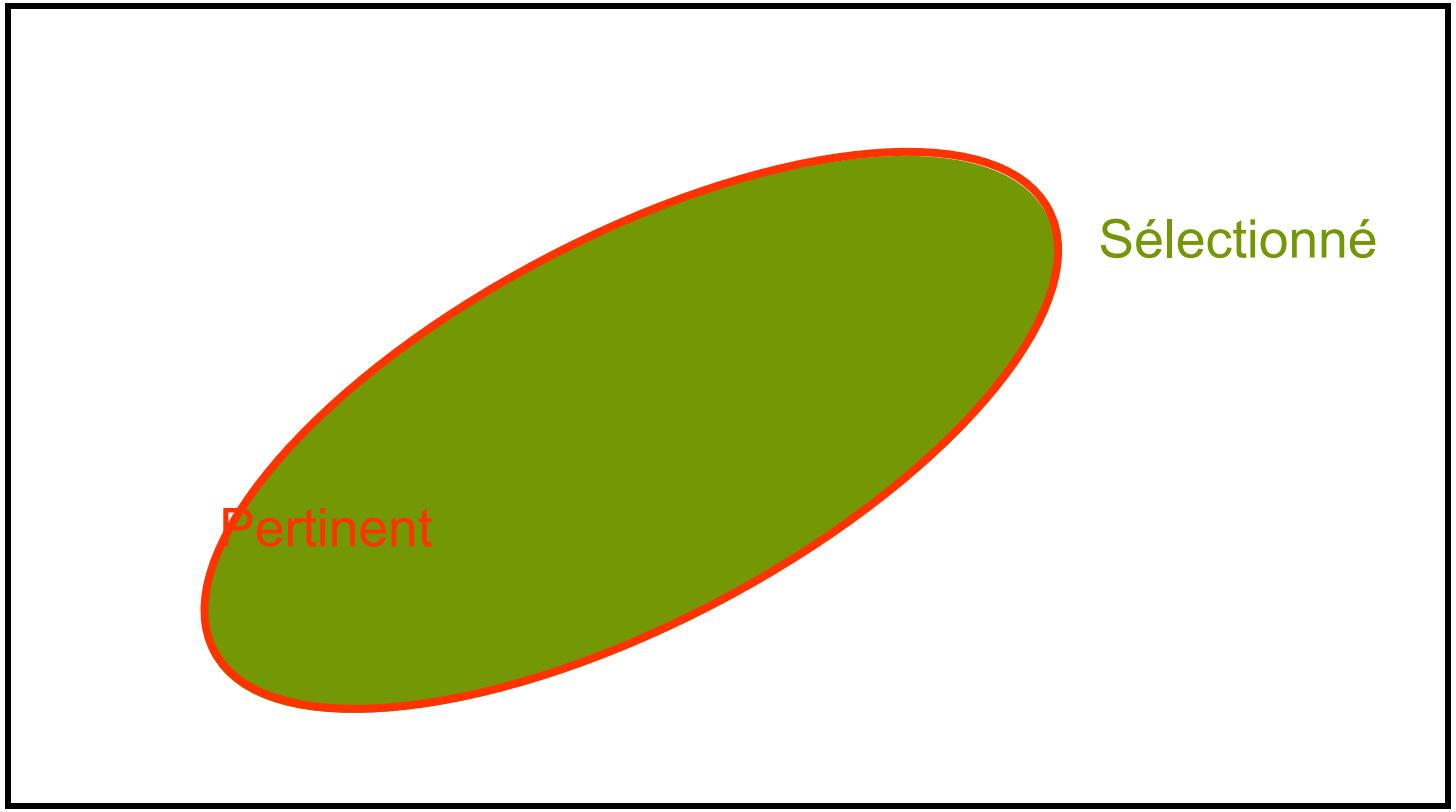
Sélectionné vs. Pertinent

Rappel élevé, mais précision faible



Sélectionné vs. Pertinent

Précision élevée, rappel élevé (idéal, mais difficile)



Rappel et la précision dans le cas d'une liste triée

Trois systèmes S1, S2 et S3 avec leurs résultats pour une requête donnée.
Le Nombre total de document pertinents pour cette requête est 6.

n	doc #	relevant
1	588	x
2	589	x

Rappel:

Précision:

n	doc #	relevant
1	576	
2	588	x
3	589	x

Rappel:

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	

Rappel:

Précision:

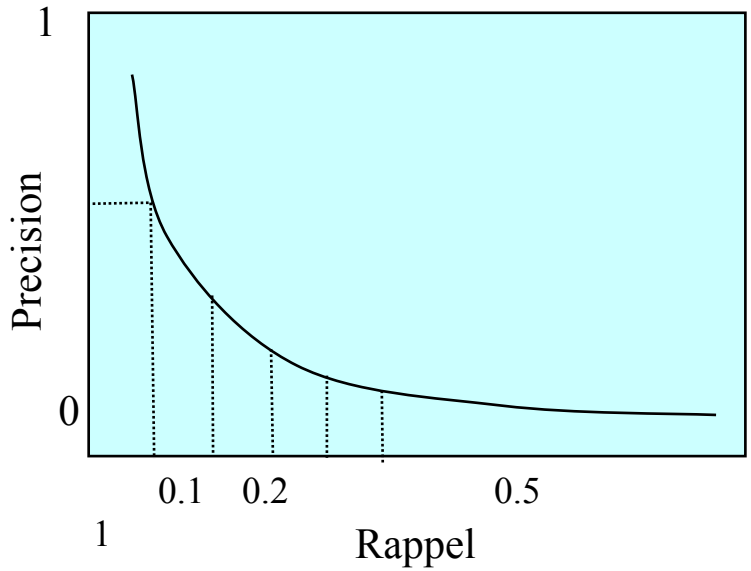
Lien entre Rappel et Précision

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 et P ?

R et P ?

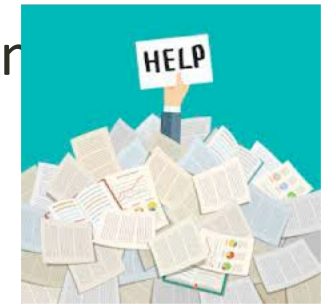
R et P ?



On calcule une Précision Moyenne (Average Precision:AP) : une seule valeur reliant le rappel et précision

Démarche d'évaluation

- Parfait, Je sais calculer le rappel et la précision sur une liste
- Hypothèse :
  - J'ai une collection de documents (exemple le Web)
  - J'ai une liste de requêtes, combien : 10?, 20? , 50?, 100?
- MAIS
  - comment savoir si un document est pertinent → je peux le faire moi-même
  - … comment connaître/identifier tous les documer Pertinents à une requête ?



Démarche d'évaluation

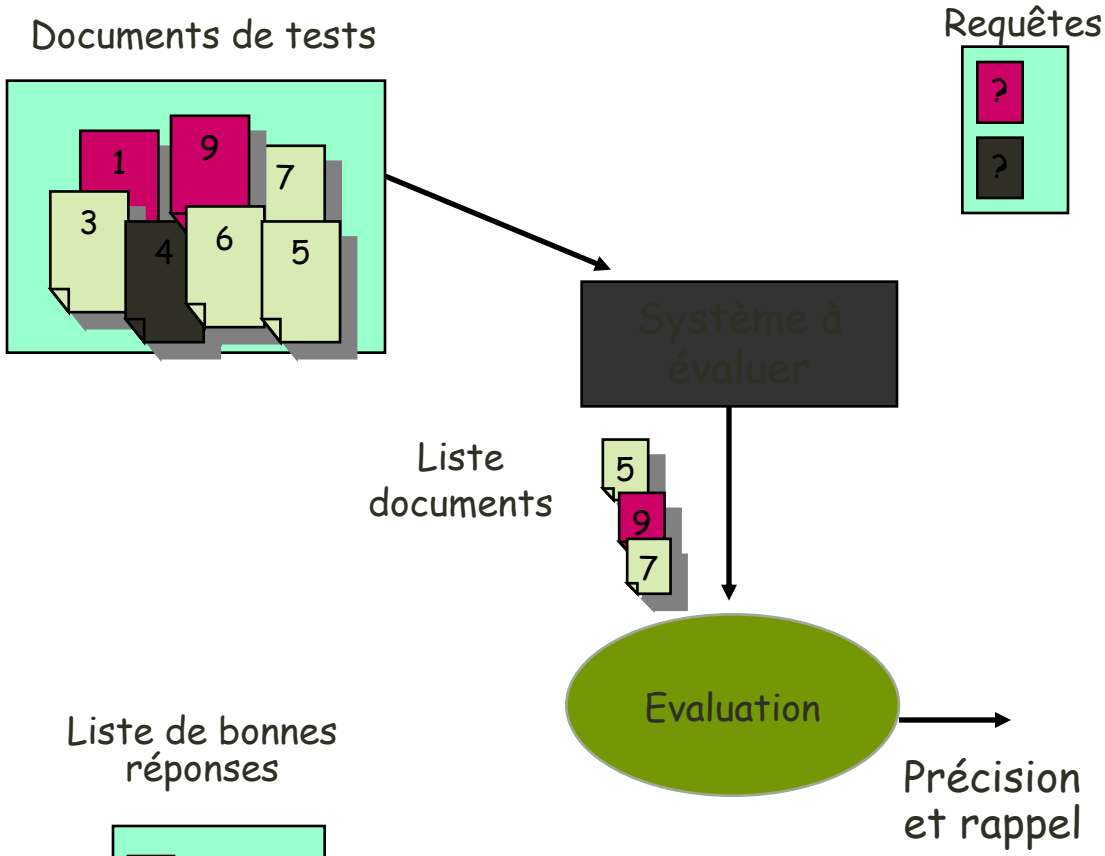
- Démarche Expérimentale (lab-based evaluation) (Cranfield Paradigm)
  - « benchmarking ».
  - Evaluation effectuée sur des collections de tests
  - Collection de test : un ensemble de documents, un ensemble de requêtes et des pertinences (réponses positives pour chaque requête)

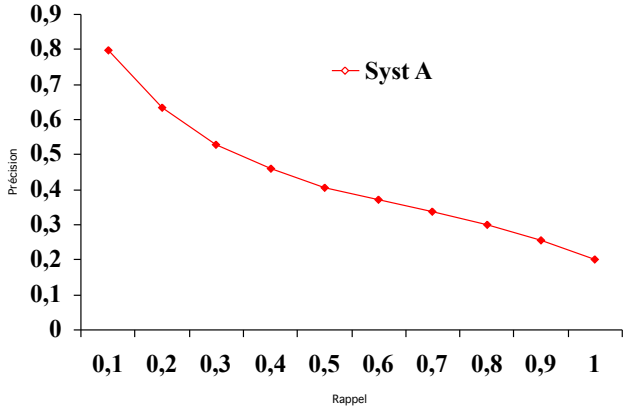
- User studies evaluation
  - RI interactive, comportement de l'utilisateur

Campagnes d'évaluation

- TREC - Text REtrieval Conference
  - Évaluation des approches RI (beaucoup de tâches sont évaluées dans cette campagne)
- CLEF - Cross Language Evaluation Forum
  - Évaluation des approches de croisement de langues (multilinguisme)
- INEX - Initiative for the Evaluation of XML Retrieval
  - Évaluation de la RI sur des documents de type XML
- NTCIR- NII Testbeds and community for information access Research

Evaluation à la Cranfield





Calcul du rappel et la précision

Calcul du rappel et de la précision

- On suppose qu'on dispose d'une collection de test
  - Lancer chaque requête sur la collection de test
  - Marquer les documents pertinents par rapport à la liste de test.
  - Calculer le rappel et la précision à pour chaque document pertinent de la liste.

Calcul du rappel et de la précision
Exemple

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	

Le nombre total de documents pertinents est = 6

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

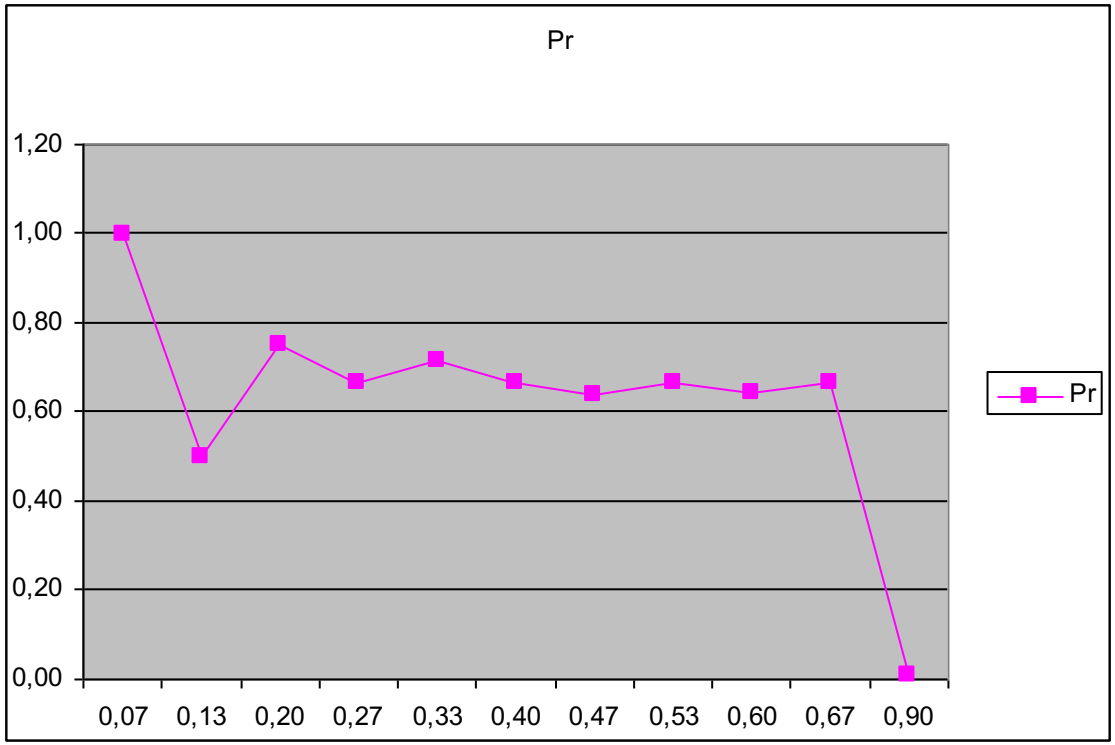
R=4/6=0.667; P=4/6=0.667

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

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

Calcul du rappel et de la précision
Exemple 2

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 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 $r>=r_j$

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

Exemple Interpolation des Précisions

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

Ra	Pr
0,0	
0,1	
0,2	
0,3	
0,4	
0,5	
0,6	
0,7	
0,8	
0,9	
1	

Précision moyenne

- On souhaite souvent avoir une valeur unique
  - Par exemple pour les algorithmes d'apprentissage pour contrôler l'amélioration
- La précision moyenne est souvent utilisée en RI
- Plusieurs moyennes
  - Précision moyenne non interpolée (PrecAvg) :
    - Calculer la moyenne des précisions à chaque apparition d'un document pertinent

Précision moyenne non interpolée
Exemple

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	

Le nombre total de document pertinent est = **6**

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

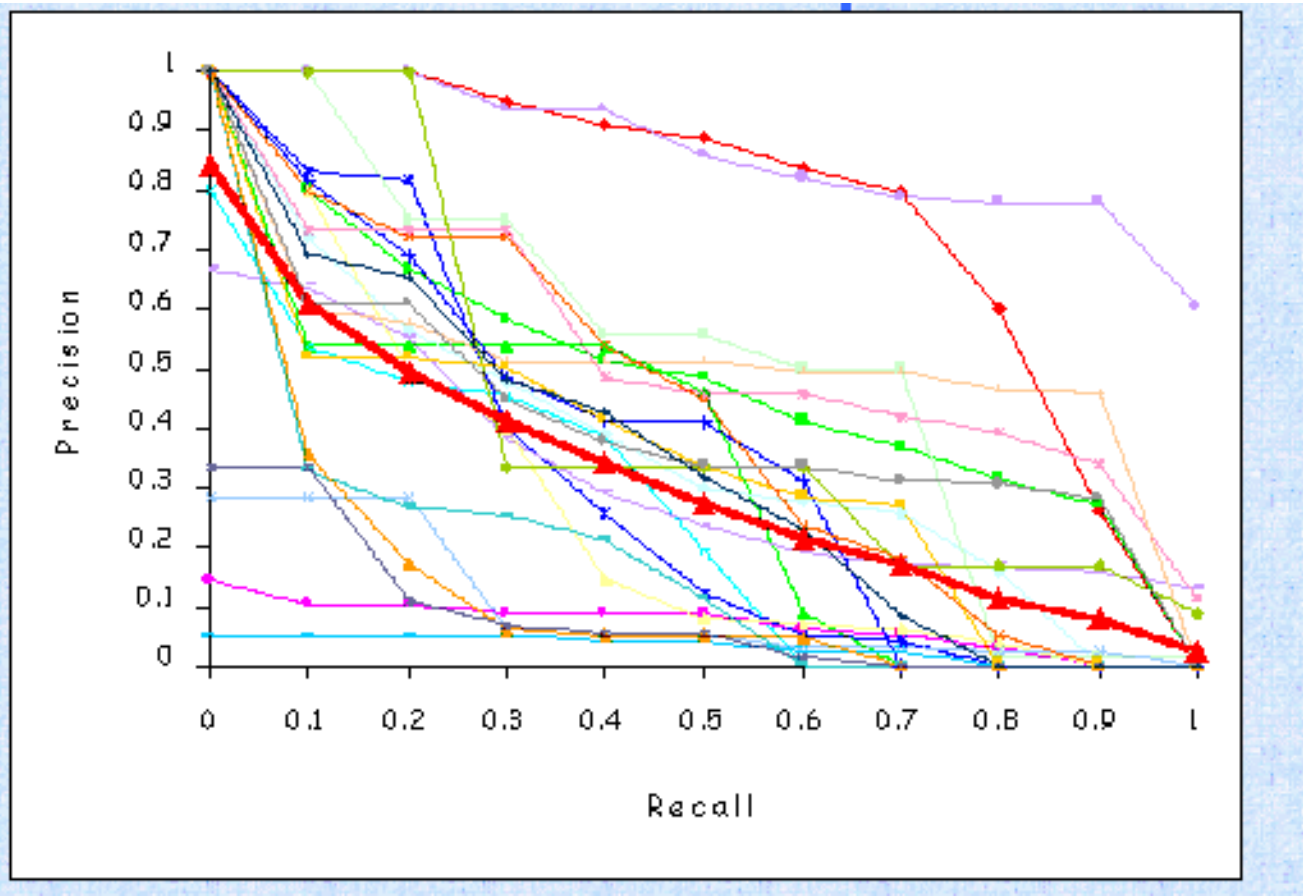
R=5/6=0.833; p=5/13=0.38

AP¨=AvgPrec=(1+1+0,75+0,667+0,38)/**6**

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

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

Exemple

Requete1	
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

AP non interpollée	**0,2329**

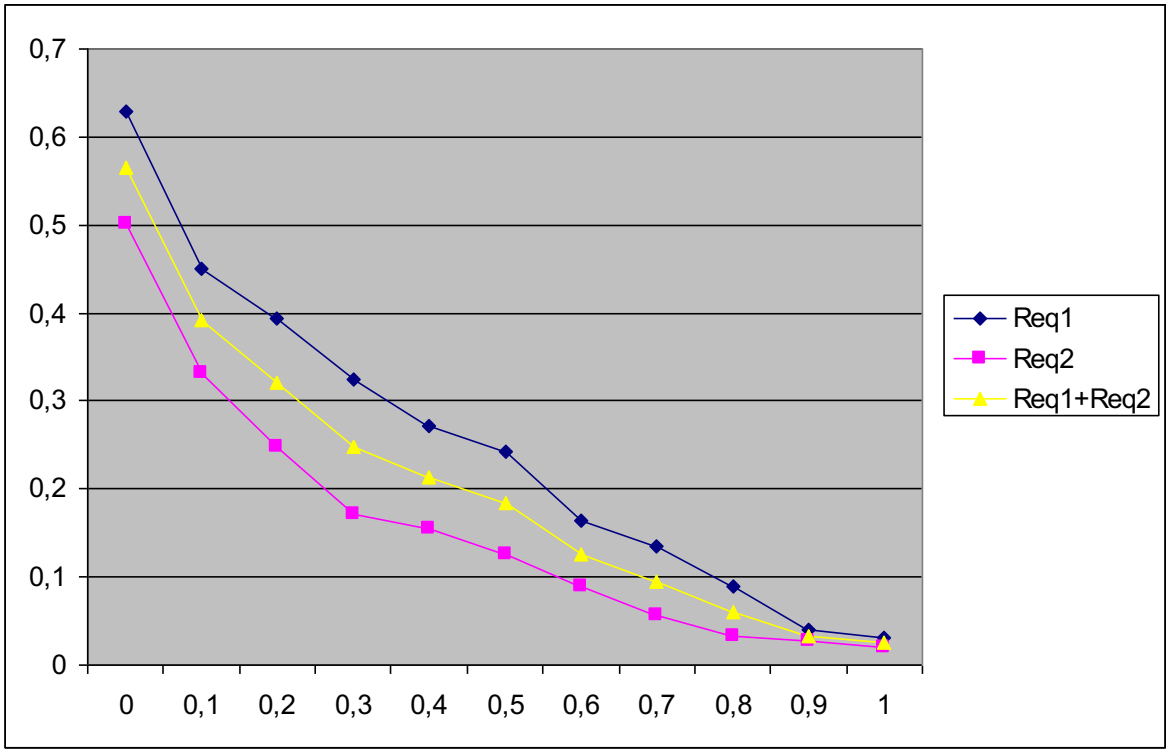
Requete2	
R	**Pr**
0	0,5017
0,1	0,332
0,2	0,248
0,3	0,171
0,4	0,155
0,5	0,125
0,6	0,089
0,7	0,056
0,8	0,032
0,9	0,027
1	0,02

AP	**0,1443**

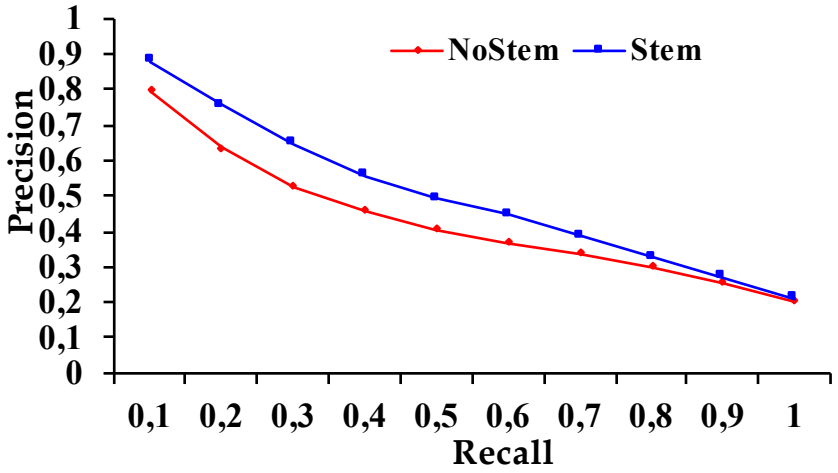
Ens des requêtes	
R	**Pr**
0	0,56535
0,1	0,3915
0,2	0,3205
0,3	0,24765
0,4	0,213
0,5	0,1837
0,6	0,1265
0,7	0,095
0,8	0,061
0,9	0,0335
1	0,0255

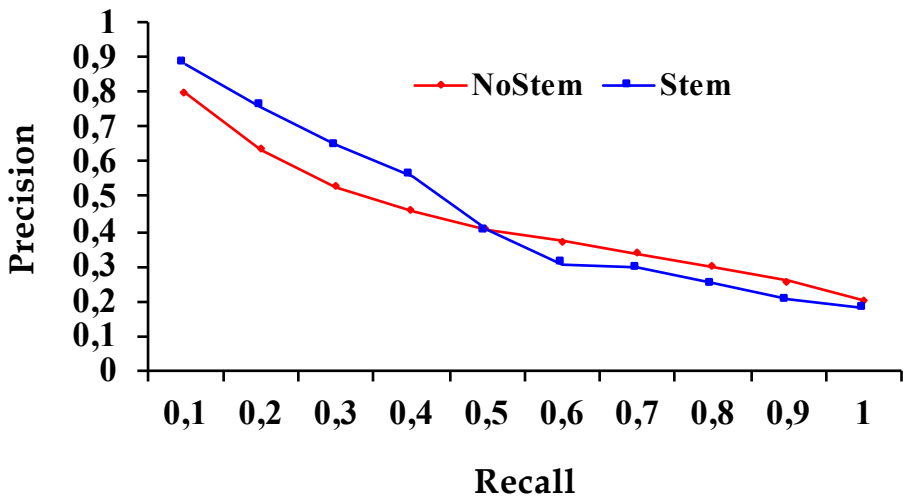
MAP	**0,1886**

Exemple


0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1
Req1
Req2
Req1+Req2

Comparaison de deux systèmes sur un ensemble de requêtes





Mesures focalisées sur le "*top*" de la liste

- Les utilisateurs se focalisent davantage sur les documents pertinents se trouvant en "top" des résultats
- La mesure de rappel n'est pas toujours appropriée
  - Il existe des stratégies de recherche pour lesquelles il y a une réponse unique
  - e.g., navigational search, question answering
- Solution : mesurer plutôt la capacité d'un SRI à trouver les documents pertinents en top de la liste

Mesures focalisées sur le “top” de la liste

- Precision au Rang X (Precision at rank X)
  - X =5, 10, 20
- Discounted Cumulative Gain
  - Prise en compte de la pertinence graduelle des documents
  - Les documents très pertinents sont plus utiles que ceux qui sont marginalement pertinents
- Reciprocal Rank
  - Rang inverse du premier document pertinent sélectionné

Précision à X documents

- Précision à différent niveau de documents
  - Précision calculée à 5 docs, 10 docs, 15docs, …

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	

Prec. à 5 docs = 3/5
Prec. à 10 docs = 4/10

R- Précision

- Une façon de calculer une valeur de précision unique : précision au **R**$^{\text{ième}}$ document de la liste des documents sélectionnés par la requête ayant **R** documents pertinents dans la collection.

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	

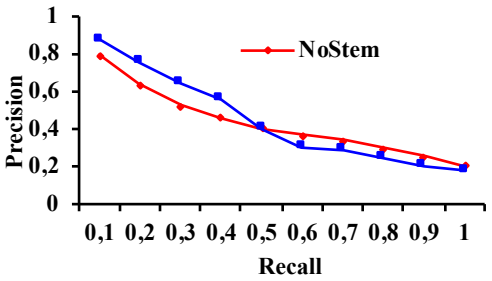
R = # documents pertinents = 6

R-Precision = 4/**6** = 0,66

Exemple *Top* X et R-précision

	Précision
at 5 docs	0,224
at 10 docs	0,177
at 15 docs	0,142
at 30 docs	0,114
at 100 docs	0,073
at 200 docs	0,053
at 500 docs	0,013
R-précision= Précision Exacte	0,144

Retour sur la Comparaison de deux systèmes



	Précision
at 5 docs	0,224
at 10 docs	0,177
at 15 docs	0,142
at 30 docs	0,114
at 100 docs	0,073
at 200 docs	0,053
at 500 docs	0,013
R-précision= Précision Exacte	0,144

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

Quelle métrique : Courbe R-P, Précision Moyenne, R-Précision, Top X, ??

Récapitulatif des métriques en RI

- R-Précision,
- MAP,
- P@X,
- RR (Reciprocal Rank)
- NDGC,
- BPREF,
- E-mesure,
- Coverage,
- Novelty.