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

Objectif

- Evaluer la performance d'une approche, d'une technique, d'un système
  - En RI, on ne mesure pas la performance absolue d'un système/technique/approche car non significative
  - Mais, ..
    - Evaluation comparative entre approches
    - Mesurer la performance relative de A par rapport à B

Critères d'évaluation

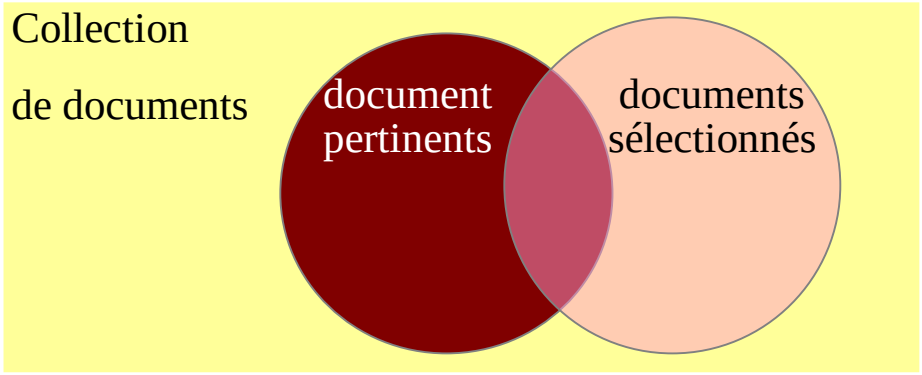
- Plusieurs critères (Cleverdon 66)
  - Facilité d'utilisation du système
  - Coût accès/stockage
  - Présentation des résultats

  - Capacité d'un système à sélectionner des documents pertinents.

Deux facteurs

- 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

Précision et Rappel



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

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

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

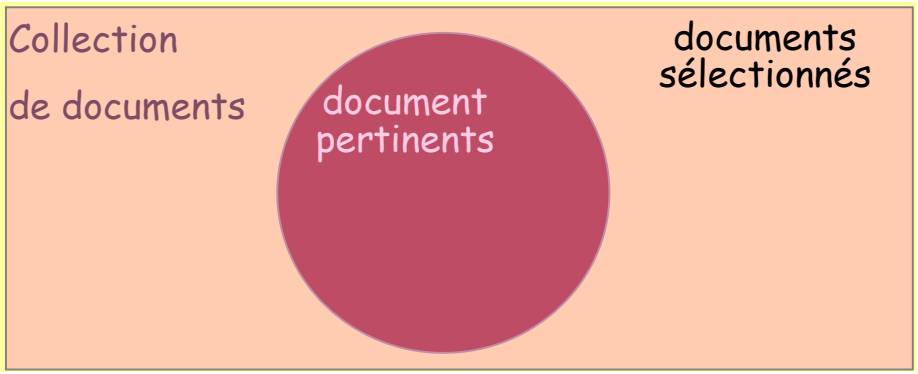
Exercice :

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

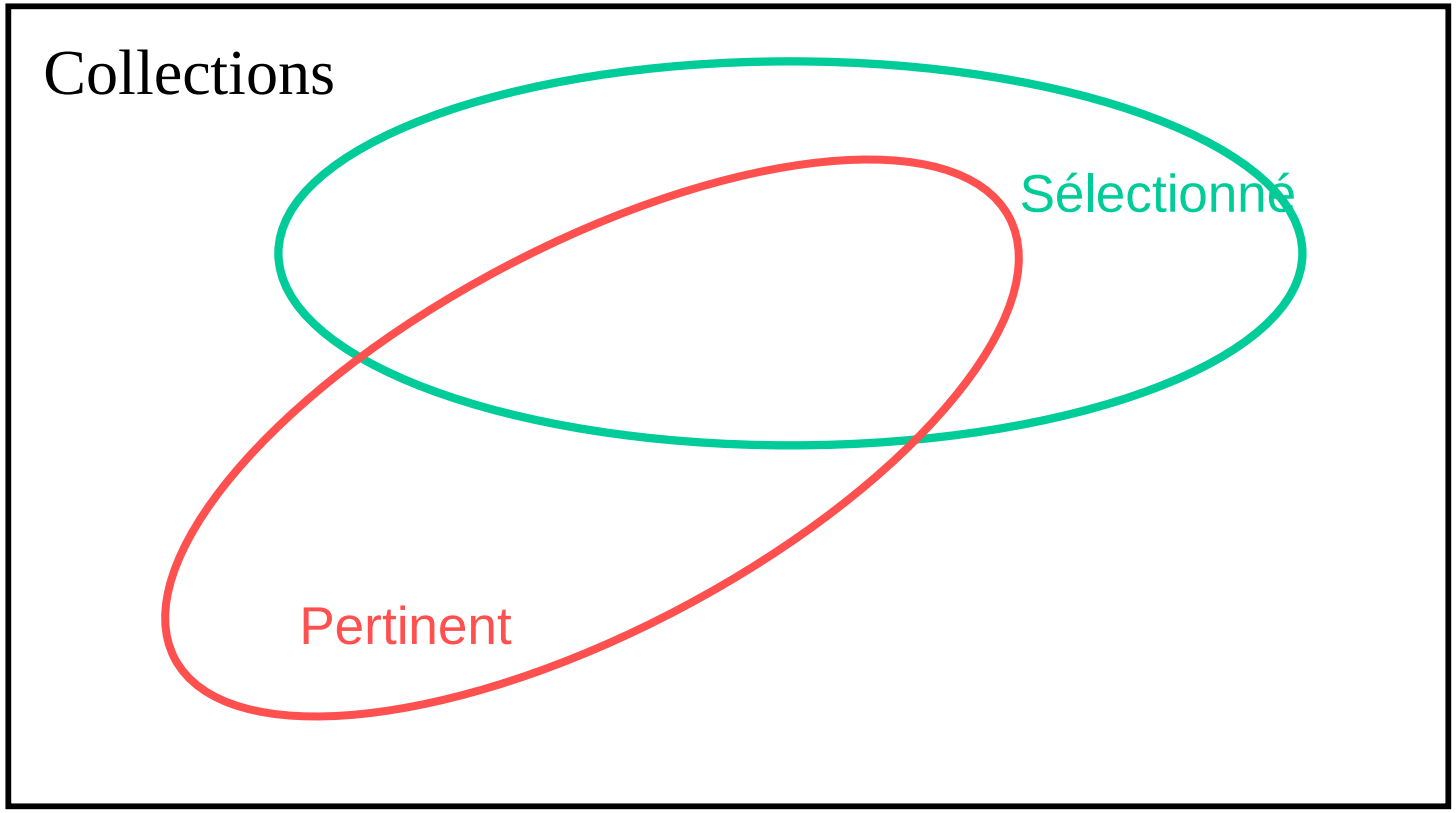
Pourquoi deux facteurs ?



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

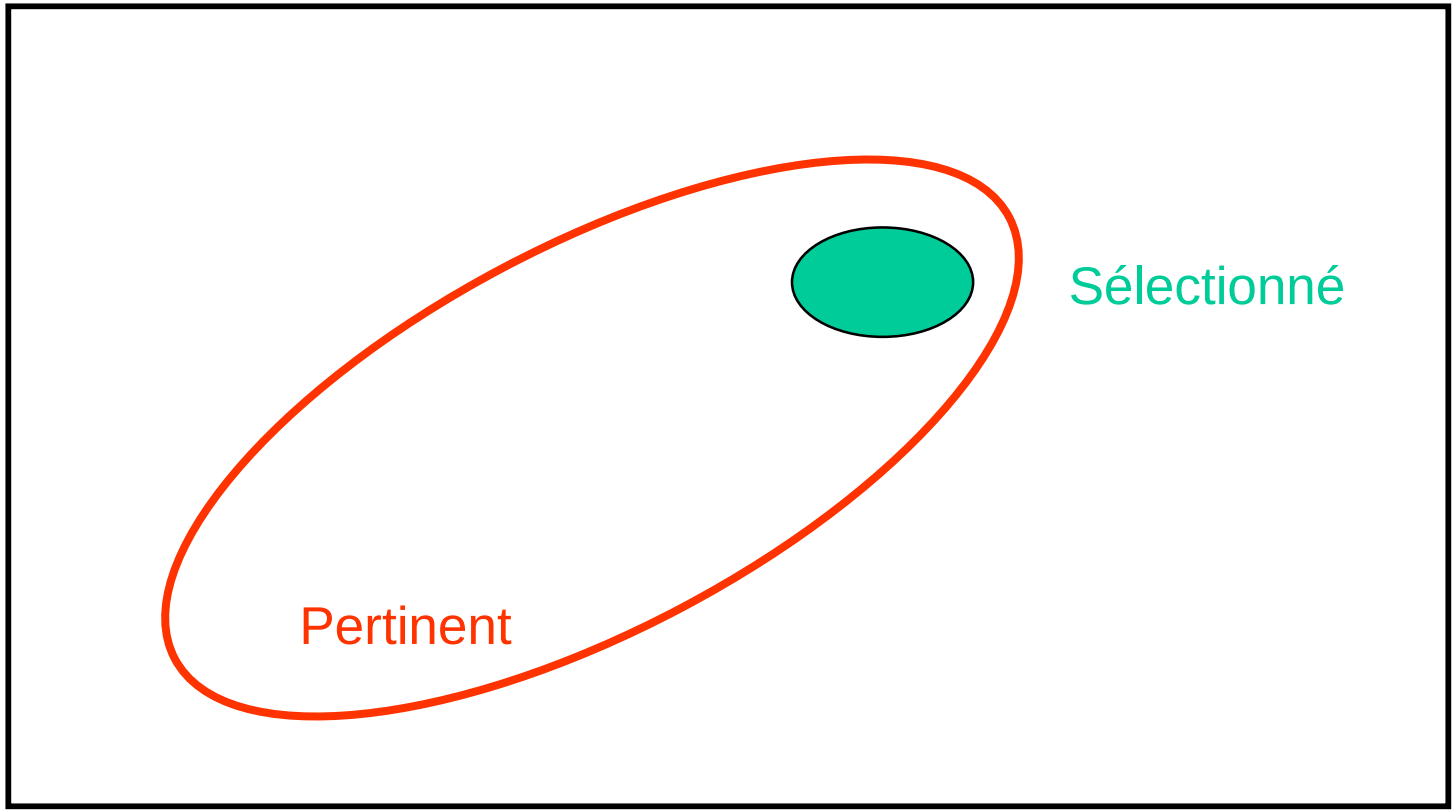
- MAIS, la précision sera très faible

Pertinent vs. Sélectionné



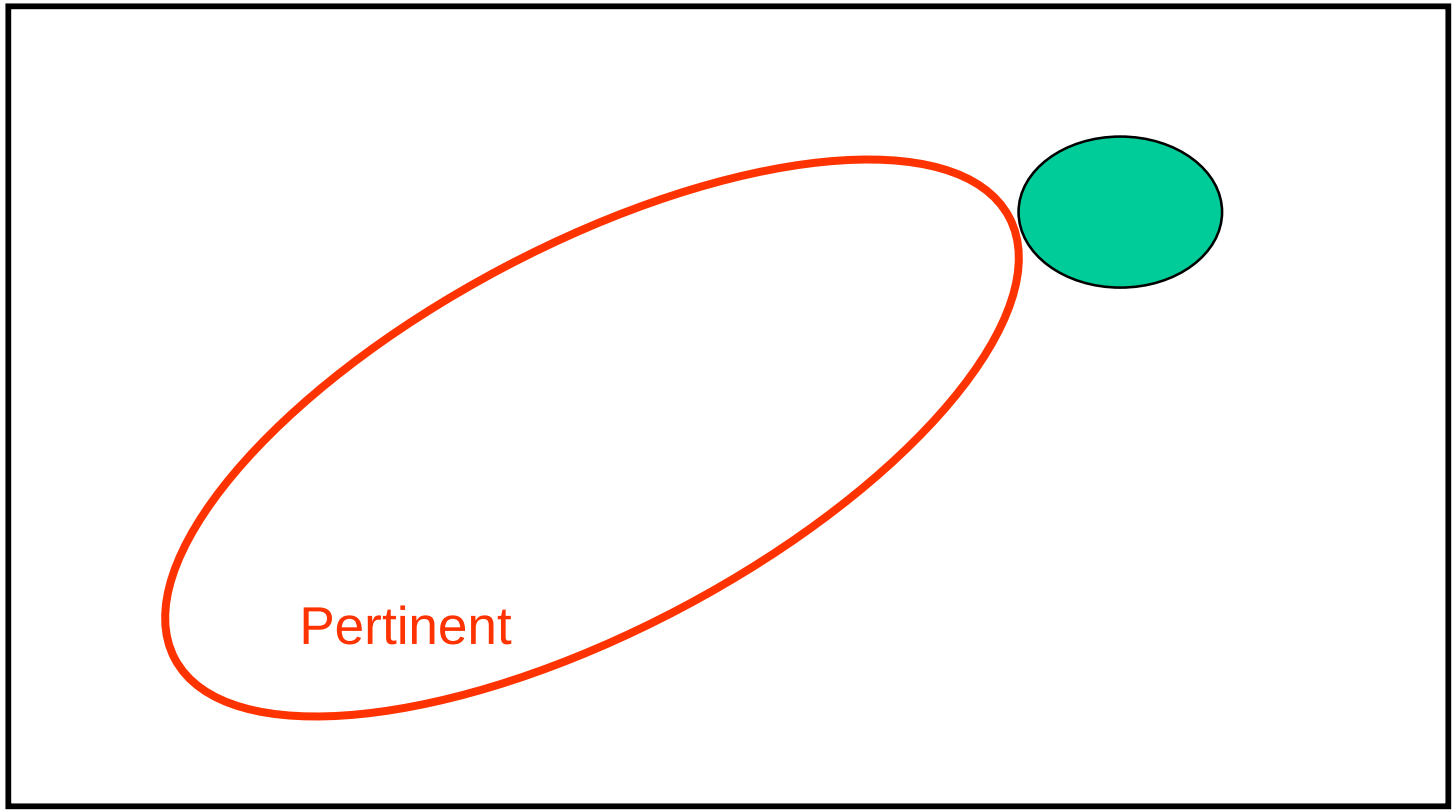
Sélectionné vs. Pertinent

Précision très élevée, rappel très faible



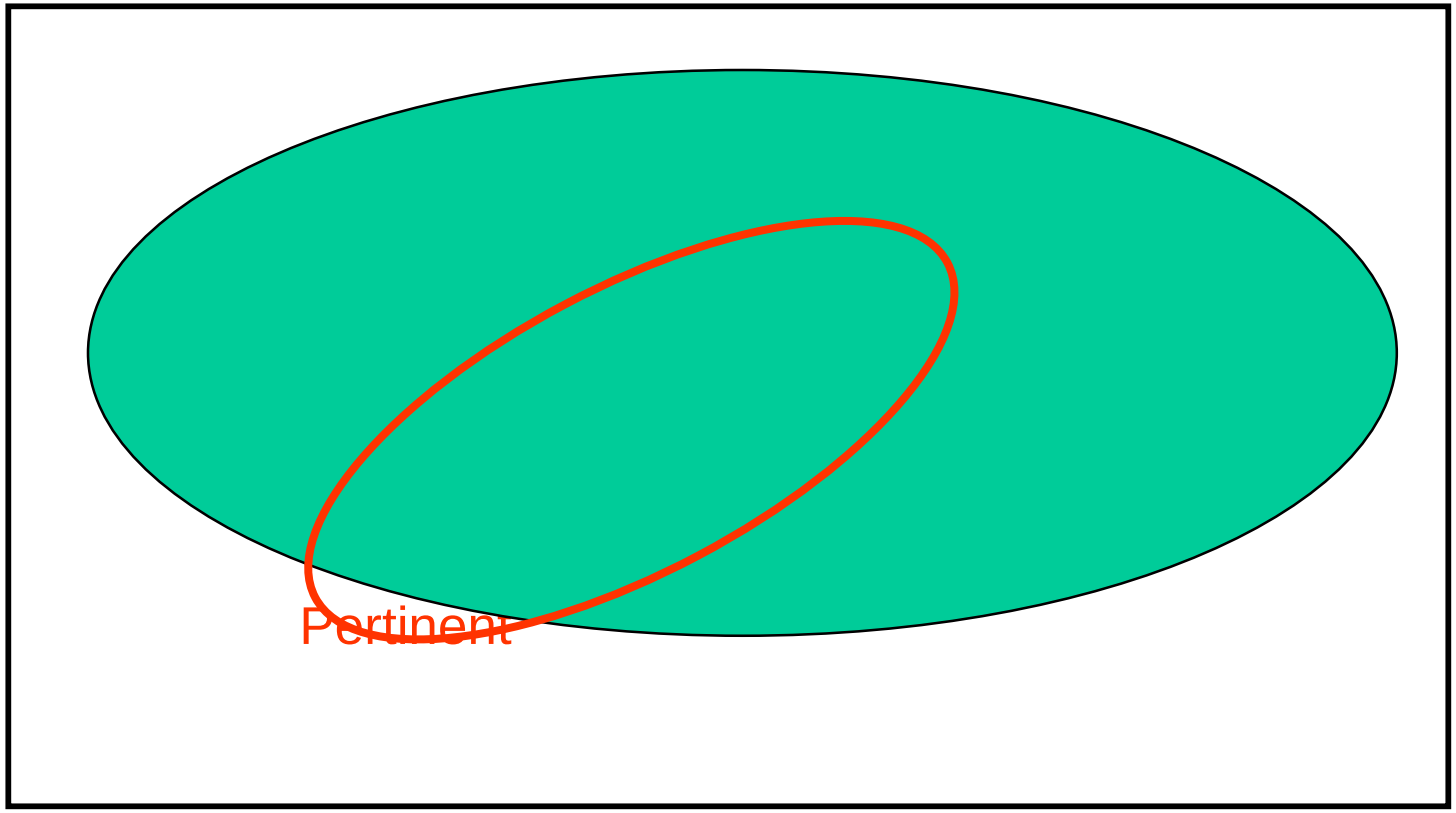
Sélectionné vs. Pertinent

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



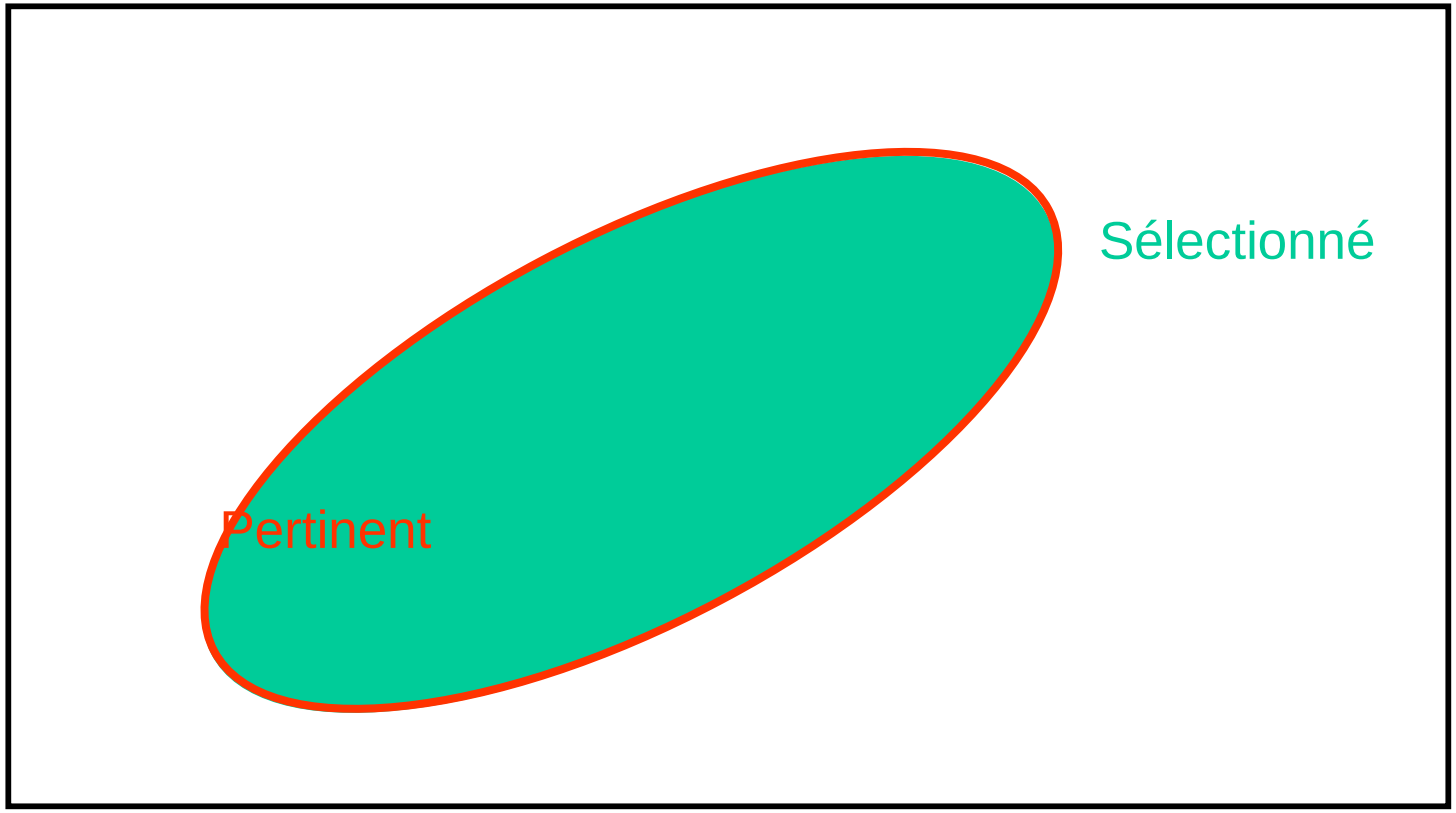
Sélectionné vs. Pertinent

Rappel élevé, mais précision faible

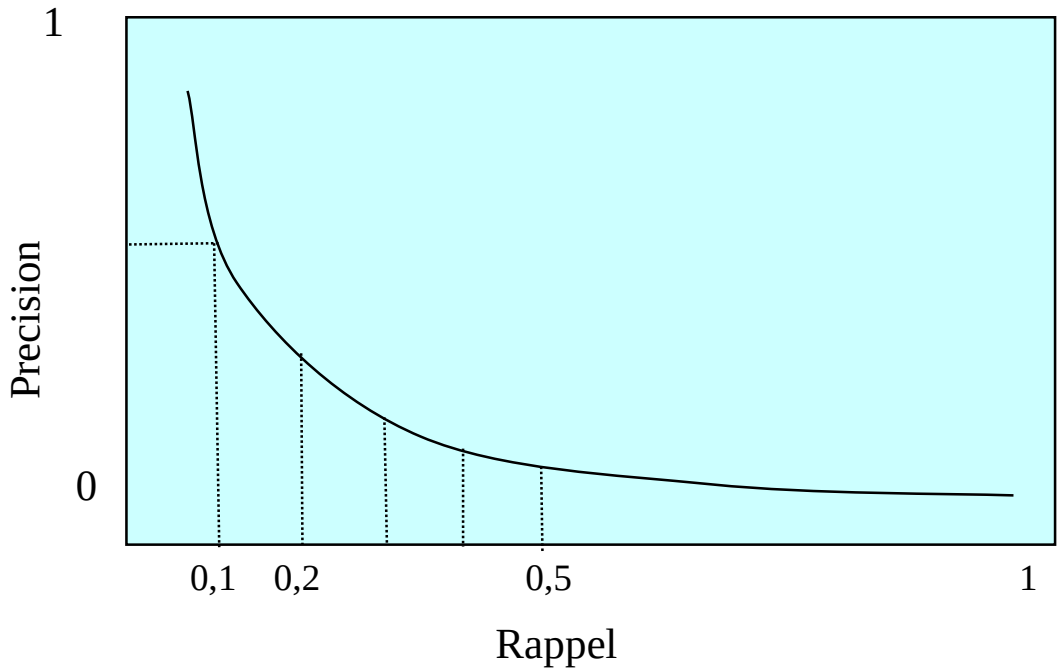


Sélectionné vs. Pertinent

Précision élevée, rappel élevé (idéal, mais difficile)



Lien entre Rappel et Précision



Précision moyenne : une seule valeur reliant le rappel et précision

Démarche d'évaluation

- **Démarche Analytique (formelle)** :
  - Difficile pour les SRI, car plusieurs facteurs : pertinence, distribution des termes, etc. sont difficiles à formaliser mathématiquement

- **Démarche Expérimentale**
  - par « **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 requêtes)

Démarche expérimentale

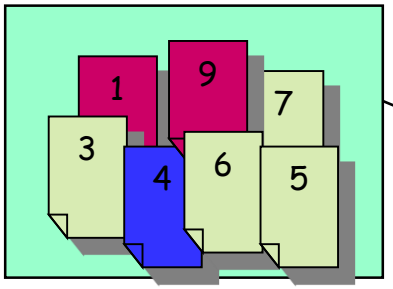
- Lancée dès les années 1960, par Cleverdon, dans le cadre du projet Cranfield

- Objectif du projet Cranfield
  - Construire des collections de test
  - Evaluer les systèmes sur ces collections de test

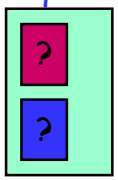
- Evaluation à la Cranfield

Evaluation à la Cranfield

Documents de tests



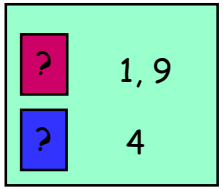
Requêtes de test



Système à évaluer

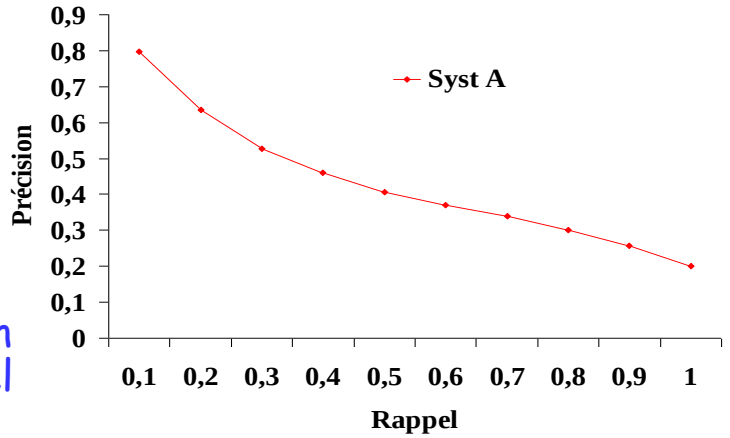
Liste documents

Liste de bonnes réponses



Evaluation

Précision et rappel



Test Collections

-

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

- COLLECTION TREC

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

Calcul du rappel et de la précision

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

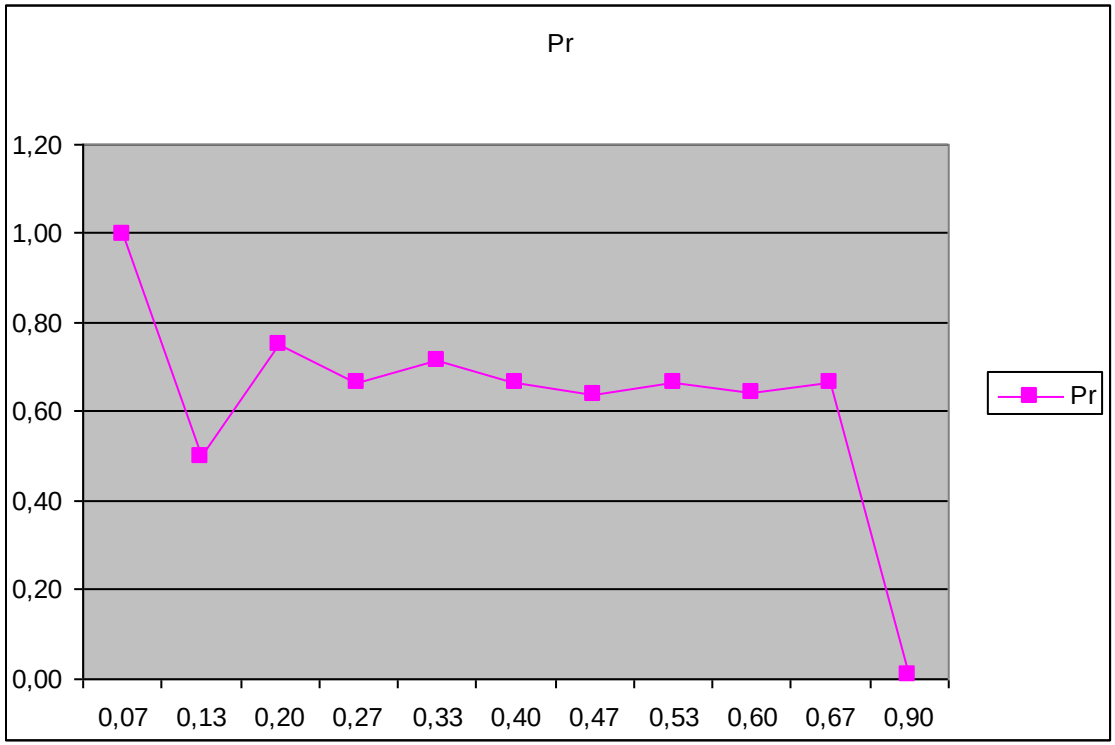
R=4/6=0.667; P=4/6=0.667

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

Il manque un document pertinent. On 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

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

Précision moyenne non interpolée pour une requête

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

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

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

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

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

Autres mesures de moyennes

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

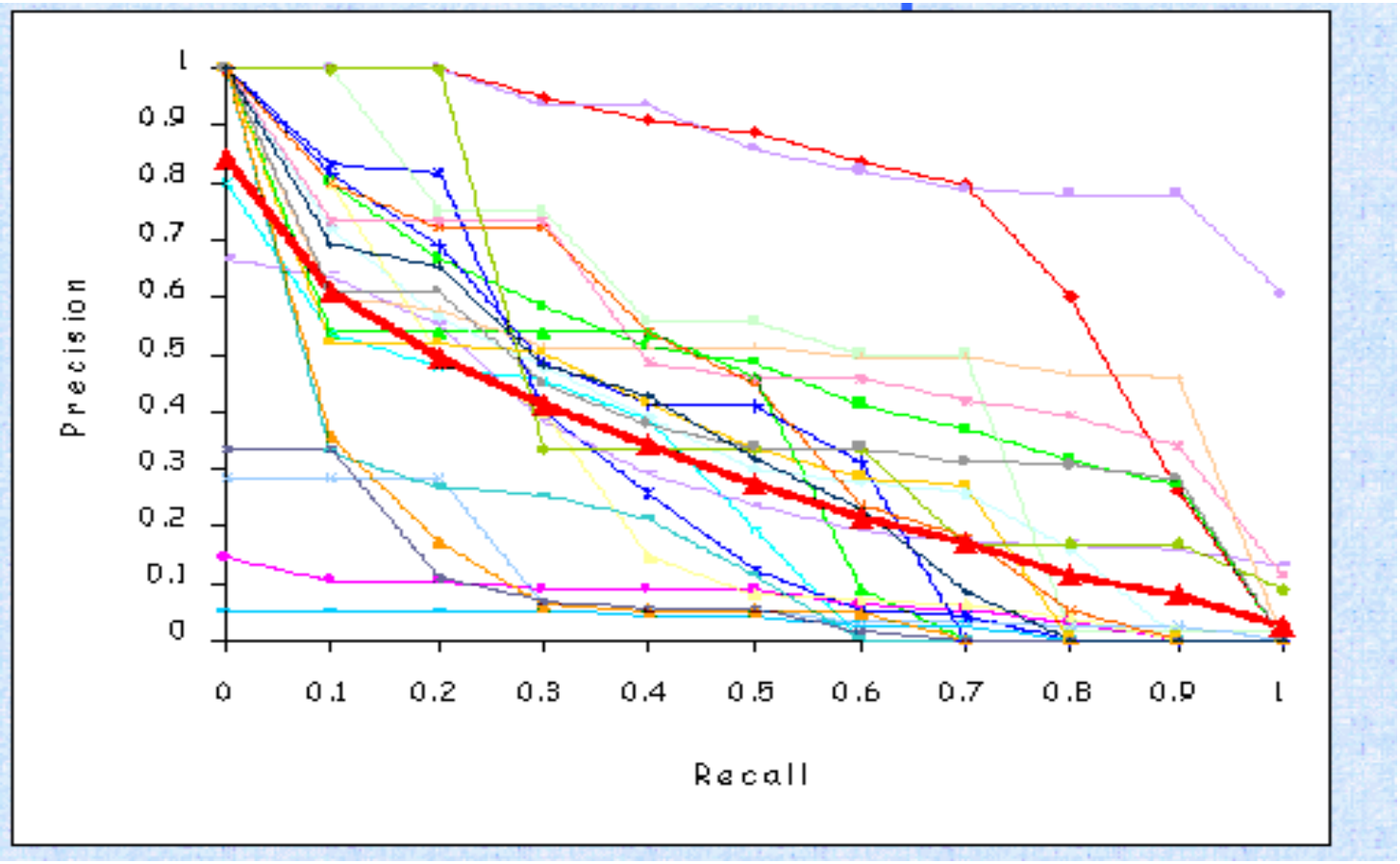
$$E = \frac{(1+\beta^2)PR}{\beta^2 P + R} = \frac{(1+\beta^2)}{\frac{\beta^2}{R}+\frac{1}{P}}$$

- β contrôle le compromis R, P:
  - β = 1: même poids précision et recall (E=F).
  - β > 1: préviligie la précision au rappel
  - β < 1: plus d'importance au rappel.

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

Moyenne sur plusieurs requêtes

- Deux façons de calculer la moyenne
  - Micro-moyenne – chaque document pertinent est un point de la moyenne
  - Macro-moyenne – faire la moyenne par requête

- On calcule également la moyenne des précisions moyennes

Courbe des moyennes sur plusieurs requêtes

- Macro moyenne
  - Calculer la précision à chaque point de rappel (précision interpolée) pour l'ensemble des requêtes.
  - Tracer la courbe rappel-précision

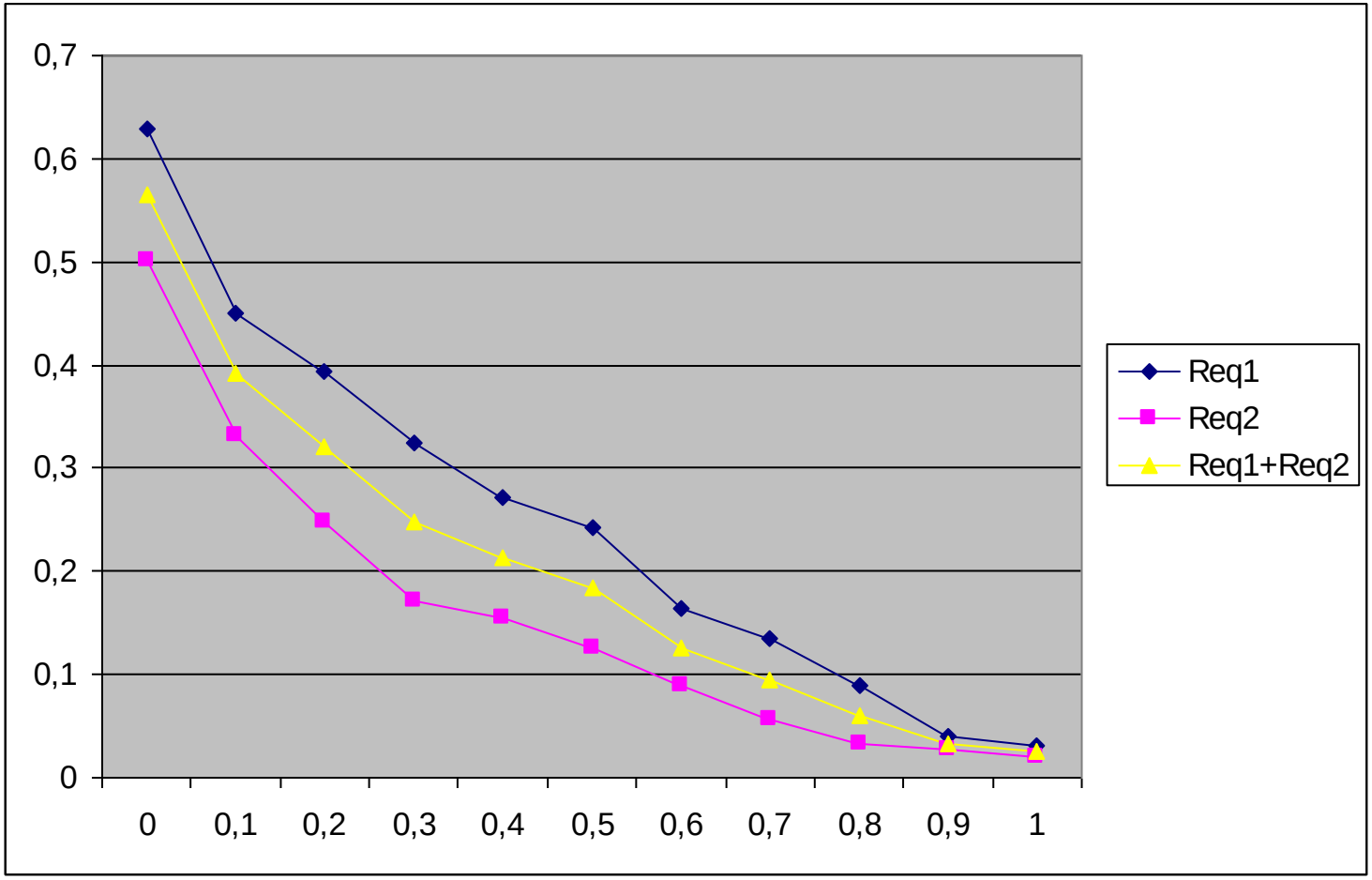
Exemple

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
AvrPrec	**0,2329**

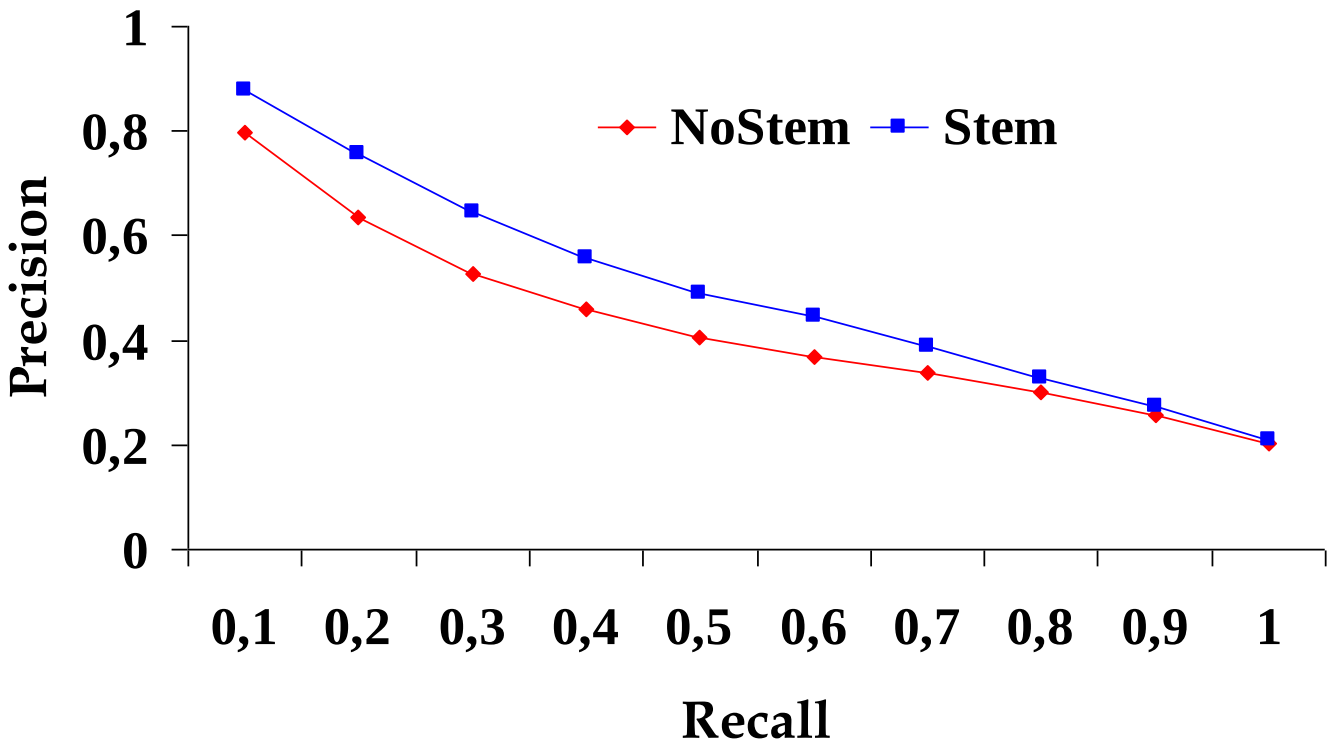
Requete2	
R	**Pr**
0	0,5017
0,1	0,332
0,2	0,248
0,3	0,171
0,4	0,155
0,5	0,125
0,6	0,089
0,7	0,056
0,8	0,032
0,9	0,027
1	0,02
AvrPrec	**0,1443**

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
AvrPrec	**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 ème document de la liste des documents sélectionné 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 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 Exact	0,144

Discounted Cumulative Gain

- Deux hypothèses:
  - Prise en compte de la pertinence graduelle des documents
  - Les documents très pertinents sont plus utiles que ceux qui sont marginalement pertinents
  - Plus un document pertinent est loin du début de la liste moins il est utile pour l'utilisateur, car il a peu de chance d'être examiné

Discounted Cumulative Gain

- Utilise la pertinence graduelle comme une mesure de l'utilité, ou du gain, obtenu en examinant un document

- Le gain est accumulé en commençant par le haut du classement et il est réduit (diminué) au fur et à mesure on l'on va vers le fond de la liste

- La réduction peut être de 1/*log (rang)*
  - Avec un log base 2, une réduction au rang 4 serait de 1/2, au rang 8 de 1/3

Discounted Cumulative Gain

- Le gain cumulé au rang *p*

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

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

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

- Autre formulation

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log(1+i)}$$

DCG Exemple

- Soit une liste de 10 documents jugés sur une échelle de 4 : 0, 1, 2, 3:

  3, 2, 3, 0, 0, 1, 2, 2, 3, 0

- Gain réduit (DG) = rel/$\log_2$(rank)

  3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0

  = 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0

- DCG:

  3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

DCG normalisé

- Moyenne des DCG sur un ensemble de requêtes
  - e.g., DCG au 5 est 6.89 et 9.61 rang 10

- Les valeurs de DCG sont souvent normalisées *selon la valeur DGC du classement parfait (Ideal DCG)*

NDCG Exemple

- Classement parfait:

  3, 3, 3, 2, 2, 2, 1, 0, 0, 0

- DCG idéal :

  3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 10.88, 10.88, 10.88

- NDCG (valeurs de DCG normalisées)
  - $NDCG_p = DCGp / iDCG_p$
  - 1, 0.83, 0.87, 0.76, 0.71, 0.69, 0.73, 0.8, 0.88, 0.88
  - $NDCG \leq 1$

Retour sur la comparaison de systèmes

- L'évaluation en RI est comparative
  - Vérifier si le système A est meilleur que le système B ?
  - Quelle est la démarche ?
    - Comparer les performances en termes de (précisions moyennes, R, F) des deux systèmes
      - (Val(A)-Val(B)/Val(B))*100
      - …. partir de 5% on peut considérer que A et meilleur que B
    - Comparer les courbes R/P
      - La courbe de A est toujours supérieure à celle de B
- Que se passe t-il quand on change de collection?

Tests statistiques

- The **t-test** is the standard statistical test for comparing two table of numbers, but depends on statistical assumptions of independence and normal distributions that do not apply to this data.

- The **sign test** makes no assumptions of normality and uses only the sign (not the magnitude) of the differences in the sample values, but assumes independent samples.

- The **Wilcoxon signed rank** uses the ranks of the differences, not their magnitudes, and makes no assumption of normality but but assumes independent samples.

- The **Kendall tau**.

Questions

- Comment construire une collection de test ?
  - Quels / combien de documents ?
  - Quelles / combien de requêtes ?
  - Comment identifier les documents pertinents pour chaque requête ?

- Evaluer la validité de la collection

Comment identifier les documents pertinents ?

- Pour répondre d'une façon sûre, il faut
  - Juger tous les documents de la collection pour chaque requête
  - Qui juge ?
    - Humain : 1, 2 .. n personnes
    - Faisable pour des petites collections
    - Impossible sur des collections volumineuses
      - TREC collections ont plus d'un millions de documents

Comment identifier les documents pertinents ?

- Autres approches
  - Pooling
  - Sampling

Comment identifier les documents pertinents ?

- Pooling
  - Pour chaque requête
    - Sélectionner des documents en utilisant différents techniques
    - Juger les n meilleurs documents obtenus par chaque technique
    - La liste des documents pertinents = l'union des documents pertinents d de chaque technique
    - Sous ensemble de vrai jugement de pertinence
- Echantillonnage
  - Possible d'estimer le nombre de documents pertinents par des techniques d'échantillonnage
- Incomplète, problème ?
  - Comment doit-on traiter les documents non jugés
  - Comment ceci peut affecter les performances calculées ?

Avantages et inconvénients des collections de tests

- Avantages
  - Mesures de performances
  - Possibilité de comparaison avec d'autres travaux
- Inconvénients
  - Les résultats obtenus sont propres à la collection.
  - Ne répondent pas à toutes les tâches de RI, notamment celles orientées utilisateur

TREC

Expérience de TREC

TREC
The Text REtrieval Conference

- Competition/collaboration between IR research groups worldwide
- Run by NIST, just outside Washington DC
- Common tasks, common test materials, common measures, common evaluation procedures
- Now various similar exercises (CLEF, NCTIR etc.)

The TREC Benchmark

- TREC: Text REtrieval Conference (http://trec.nist.gov/)
  Originated from the TIPSTER program sponsored by Defense Advanced Research Projects Agency (DARPA).

- Became an annual conference in 1992, co-sponsored by the National Institute of Standards and Technology (NIST) and DARPA.

- Uniform, appropriate scoring procedures

The TREC Objective

- TREC is a mordern example of the Cranfield Tradition
  - System evaluation based on text collections
- Sharing of resources and experiences in developing the benchmark.
  - With major sponsorship from government to develop large benchmark collections.
- Encourage participation from industry and academia.
- Development of new evaluation techniques, particularly for new applications.
  - Retrieval, routing/filtering, non-English collection, web-based collection, question answering.

A Brief History of TREC

- 1992: first TREC conference
  - started by Donna Harman and Charles Wayne as 1 of 3 evaluations in DARPA's TIPSTER program
  - first 3 CDs of documents from this era, hence known as the "TIPSTER" CDs
  - open to IR groups not funded by DARPA
    - 25 groups submitted runs
  - two tasks: ad hoc retrieval, routing
    - 2GB of text, 50 topics
    - primarily an exercise in scaling up systems

A Brief History of TREC

- 1993 (TREC-2)
  - true baseline performance for main tasks
- 1994 (TREC-3)
  - initial exploration of additional tasks in TREC
- 1995 (TREC-4)
  - official beginning of TREC track structure
- 1998 (TREC-7)
  - routing dropped as a main task, though incorporated into filtering track
- 2000 (TREC-9)
  - ad hoc main task dropped; first all-track TREC

fin

Pertinence

- Quelques suppositions « fausses »
  - Pertinence binaire (oui/non)
    - Les utilisateurs ne jugent pas souvent les documents par pertinent ou non pertinent
  - Pertinence d'un seul document peut être jugée indépendamment du contexte
    - Les utilisateurs peuvent juger différemment un document selon ce qu'ils ont vu au préalable.

Using Preferences

- Two rankings described using preferences can be compared using the *Kendall tau coefficient (τ):*

$$\tau = \frac{P - Q}{P + Q}$$

  - *P* is the number of preferences that agree and *Q* is the number that disagree

- For preferences derived from binary relevance judgments, can use *BPREF*

BPREF

- For a query with *R* relevant documents, only the first *R* non-relevant documents are considered

$$BPREF = \frac{1}{R} \sum_{d_r} (1 - \frac{N_{d_r}}{R})$$

  - d_r is a relevant document, and N_{dr} gives the number of non-relevant documents

- Alternative definition

$$BPREF = \frac{P}{P+Q}$$

Efficiency Metrics

Metric name	Description
Elapsed indexing time	Measures the amount of time necessary to build a document index on a particular system.
Indexing processor time	Measures the CPU seconds used in building a document index. This is similar to elapsed time, but does not count time waiting for I/O or speed gains from parallelism.
Query throughput	Number of queries processed per second.
Query latency	The amount of time a user must wait after issuing a query before receiving a response, measured in milliseconds. This can be measured using the mean, but is often more instructive when used with the median or a percentile bound.
Indexing temporary space	Amount of temporary disk space used while creating an index.
Index size	Amount of storage necessary to store the index files.

Significance Tests

- Given the results from a number of queries, how can we conclude that ranking algorithm A is better than algorithm B?
- A significance test enables us to reject the *null hypothesis* (no difference) in favor of the *alternative hypothesis* (B is better than A)
  - the *power* of a test is the probability that the test will reject the null hypothesis correctly
  - increasing the number of queries in the experiment also increases power of test

Significance Tests

1. Compute the effectiveness measure for every query for both rankings.

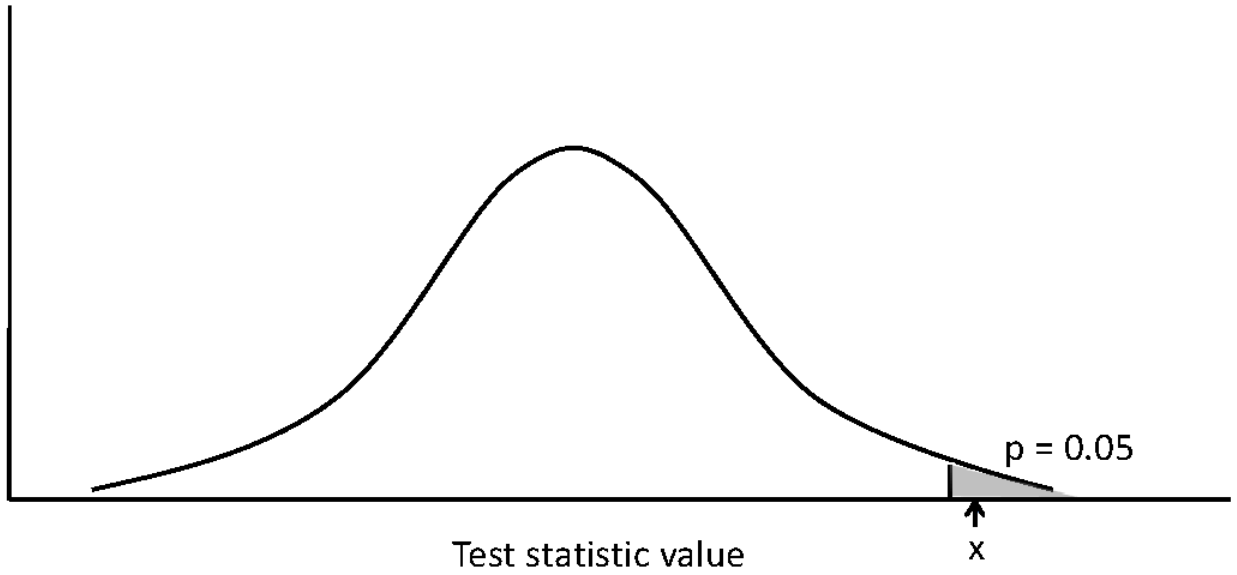
2. Compute a *test statistic* based on a comparison of the effectiveness measures for each query. The test statistic depends on the significance test, and is simply a quantity calculated from the sample data that is used to decide whether or not the null hypothesis should be rejected.

3. The test statistic is used to compute a *P-value*, which is the probability that a test statistic value at least that extreme could be observed if the null hypothesis were true. Small P-values suggest that the null hypothesis may be false.

4. The null hypothesis (no difference) is rejected in favor of the alternate hypothesis (i.e., B is more effective than A) if the P-value is $\leq \alpha$, the *significance level*. Values for α are small, typically .05 and .1, to reduce the chance of a Type I error.

One-Sided Test

- Distribution for the possible values of a test statistic assuming the null hypothesis



- shaded area is *region of rejection*

Example Experimental Results

Query	A	B	B-A
1	25	35	10
2	43	84	41
3	39	15	-24
4	75	75	0
5	43	68	25
6	15	85	70
7	20	80	60
8	52	50	-2
9	49	58	9
10	50	75	25

t-Test

- Assumption is that the difference between the effectiveness values is a sample from a normal distribution
- Null hypothesis is that the mean of the distribution of differences is zero
- Test statistic

$$t = \frac{\overline{B-A}}{\sigma_{B-A}} . \sqrt{N}$$

  - for the example,

$$\overline{B-A} = 21.4,\ \sigma_{B-A} = 29.1,\ t = 2.33,\ \text{p-value}{=}.02$$

Wilcoxon Signed-Ranks Test

- Nonparametric test based on differences between effectiveness scores
- Test statistic

$$w = \sum_{i=1}^{N} R_i$$

R_i is a signed-rank, N is the number of differences $\neq 0$

  - To compute the signed-ranks, the differences are ordered by their absolute values (increasing), and then assigned rank values
  - rank values are then given the sign of the original difference

Wilcoxon Example

- 9 non-zero differences are (in rank order of absolute value):

  2, 9, 10, 24, 25, 25, 41, 60, 70

- Signed-ranks:

  -1, +2, +3, -4, +5.5, +5.5, +7, +8, +9

- w = 35, p-value = 0.025

Sign Test

- Ignores magnitude of differences
- Null hypothesis for this test is that
  - $P(B > A) = P(A > B) = ½$
  - number of pairs where B is “better” than A would be the same as the number of pairs where A is “better” than B
- Test statistic is number of pairs where B>A
- For example data,
  - test statistic is 7, p-value = 0.17
  - cannot reject null hypothesis

Setting Parameter Values

- Retrieval models often contain parameters that must be tuned to get best performance for specific types of data and queries
- For experiments:
  - Use *training* and *test* data sets
  - If less data available, use *cross-validation* by partitioning the data into *K* subsets
  - Using training and test data avoids *overfitting* – when parameter values do not generalize well to other data

Finding Parameter Values

- Many techniques used to find optimal parameter values given training data
  - standard problem in machine learning
- In IR, often explore the space of possible parameter values by *brute force*
  - requires large number of retrieval runs with small variations in parameter values (*parameter sweep)*
- *SVM optimization* is an example of an efficient procedure for finding good parameter values with large numbers of parameters

Online Testing

- Test (or even train) using live traffic on a search engine
- Benefits:
  - real users, less biased, large amounts of test data
- Drawbacks:
  - noisy data, can degrade user experience
- Often done on small proportion (1-5%) of live traffic

Summary

- No single measure is the correct one for any application
  - choose measures appropriate for task
  - use a combination
  - shows different aspects of the system effectiveness
- Use significance tests (t-test)
- Analyze performance of individual queries

Query Summary