UNIVERSITÉ DES SCIENCES ET TECHNOLOGIE HOUARI BOUMEDIENE

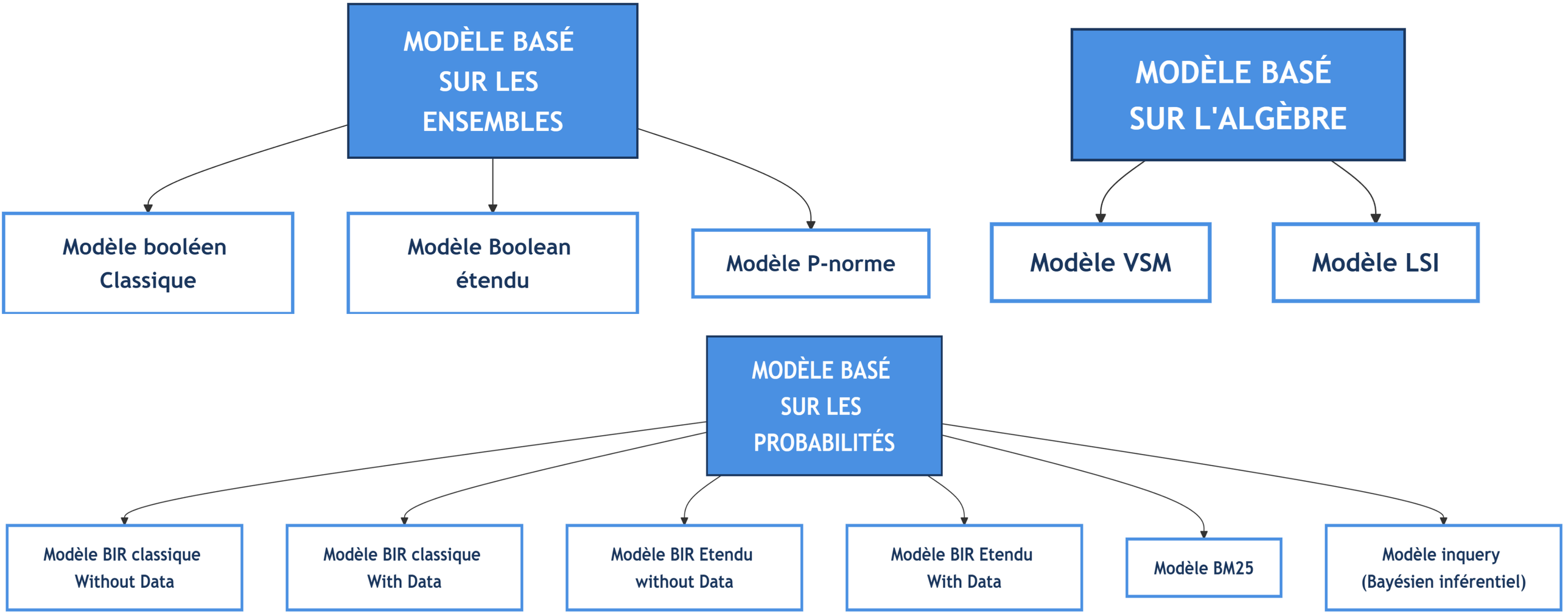
FACULTÉ D'INFORMATIQUE - MASTER 2 SII

RECHERCHE D'INFORMATION
INFORMATION RETRIEVAL

CHAPITRE 8: Evaluation des performances des Systèmes de RI

17 Novembre 2025

Les modèles de RI étudiés



I. INTRODUCTION

Pour chaque modèle, nous avons :

- calculé les RSV pour diverses requêtes,
- obtenu pour chaque modèle un classement de documents,
- observé que ces classements diffèrent selon le modèle.

Mais lequel produit réellement les “meilleurs” résultats ?

Comment dire objectivement qu’un modèle est supérieur à un autre ?

EN ÉVALUANT LES PERFORMANCES DE CHAQUE MODÈLE

II. Objectif de l'évaluation des SRI

L'objectif de cette évaluation en Recherche d'Information est la **comparaison entre des Systèmes de Recherche d'Information** (SRI).

- On ne mesure pas la performance absolue d'un SRI, car elle serait non significative.
- On mesure la performance relative d'un SRI par rapport à un autre, afin de savoir lequel est plus performant dans un contexte donné.

III. Démarche d'évaluation

III.1. Démarche analytique (formelle)

- Cette démarche consiste à évaluer un (SRI) *uniquement* à partir d'analyses théoriques ou mathématiques.
- L'idée est de déduire, par des raisonnements mathématiques, quelle approche serait « la meilleure » en supposant un modèle parfaitement formalisé.
- Cette démarche **ne peut pas être applicable en RI** pour évaluer les SRI.

Parce que la RI repose sur des phénomènes qui sont : **complexes**, **variables**, **souvent subjectifs**, **et très difficiles à modéliser**.

III. Démarche d'évaluation

III.1. Démarche analytique (formelle)

Par exemple, pour comparer deux approches, il faudrait être capable de créer des formules mathématiques fiables pour représenter :

- **L'indexation des documents** : comment les contenus sont représentés. Difficile, car les documents sont longs, variés, pleins d'ambiguïtés linguistiques.
- **La pertinence** qu'est-ce qu'un document « pertinent ». notion subjective : deux utilisateurs peuvent ne pas être d'accord.
- **La requête**.
- **La distribution des termes dans la collection.**

III. Démarche d'évaluation

III.1. Démarche analytique (formelle)

- ✓ Comme ces éléments sont **difficiles à formaliser**, on ne peut pas compter sur une approche purement mathématique pour déterminer quel système est le meilleur.
- ✓ **La méthode analytique seule ne suffit pas pour évaluer les SRI, car la réalité linguistique et cognitive ne se laisse pas capturer entièrement par des formules.**

III. Démarche d'évaluation

III.2. Démarche expérimentale (benchmarking)

- Face aux limites de la démarche analytique, la RI s'appuie sur une démarche expérimentale.
- L'idée est simple : **on teste réellement les systèmes**, on mesure leurs performances sur des jeux de données bien définis, et on compare.
- Cette démarche repose donc sur :

  ✓ un **environnement de test**,

  ✓ des **expérimentations**,

  ✓ et des **mesures d'évaluation quantitatives**

C'est aujourd'hui **la méthode dominante en RI**, parce qu'elle est flexible, objective (basée sur des chiffres), applicable à n'importe quel modèle.

III. Démarche d'évaluation

III.2. Démarche expérimentale (benchmarking)

2.2.1. Environnement de test

- Un environnement de test doit contenir au minimum :
  1. **Un ensemble de documents** (collection) : Ce sont les textes sur lesquels on va faire les recherches.
  2. **Un ensemble de requêtes de test :** Ce sont des requêtes rédigées par des experts, typiques des besoins d'un utilisateur réel.
  3. **Les documents pertinents** associés à chaque requête (jugements de pertinence) : Pour chaque requête, on indique quels documents sont pertinents.

- Ces éléments permettent de comparer les SRI dans des conditions identiques.

III. Démarche d'évaluation

III.2. Démarche expérimentale (benchmarking)

2.2.1. Environnement de test

Exemples de collections de test

Plusieurs environnements standards sont utilisés en RI depuis des décennies :

- **CACM** : articles de la revue *Communications of the ACM*.
- **CISI** : abstracts scientifiques du domaine de l'information.
- **CRAN** : résumés sur l'aéronautique.
- **MED** : documents médicaux.
- **TIME** : textes de presse.
- **TREC** : la plus grande et la plus importante collection moderne, utilisée pour la recherche depuis les années 90.

TREC est aujourd'hui le standard mondial pour tester les systèmes modernes, moteurs de recherche, modèles de langue, etc.

IV. Critères d'évaluation

Avant même de choisir une mesure, il faut comprendre **ce qu'on cherche à évaluer**. C'est une étape essentielle en RI, car la performance dépend de la *tâche*, du *type de système* et du *contexte d'usage*.

1. Identifier la tâche à évaluer

Chaque système de RI peut viser un objectif différent.

Exemples de tâches :

- Trouver les documents les plus pertinents (moteur de recherche classique).
- Récupérer tous les documents pertinents (santé, juridique).
- Trouver une réponse courte (QA, systèmes question-réponse).
- Recommander des documents (systèmes de recommandation).
- Classer ou filtrer des documents (filtrage d'information).

Selon la tâche, les critères d'évaluation ne seront pas les mêmes.

IV. Critères d'évaluation

2. Identifier les critères d'évaluation (Cleverdon 1966)

- Cleverdon est l'un des pionniers de l'évaluation en RI (projet Cranfield). Il a proposé plusieurs critères dont :
  a) **Facilité d'utilisation du système :** À quel point l'utilisateur peut-il interagir facilement avec le système.
  b) **Coût d'accès / stockage**
     - Coût de stockage de la collection
     - Temps et ressources nécessaires pour répondre aux requêtes
     - Coût matériel / logiciel
  c) **Présentation des résultats** Comment les résultats sont affichés et organisés :
     - ordre de classement
     - regroupement par thèmes
     - mise en évidence des termes de la requête

IV. Critères d’évaluation

2. Identifier les critères d’évaluation (Cleverdon 1966)

d) Capacité du système à sélectionner des documents pertinents C’est le critère central de la RI “classique”. Il évalue la qualité du *ranking* produit par le système.

C’est ici que les deux mesures fondamentales interviennent : **le rappel** et **la précision**. Ces deux mesures ont été introduites par Cleverdon et sont encore aujourd’hui les **fondations** de presque toutes les métriques d’evaluation modernes.

IV. Critères d'évaluation

IV.1 le rappel et la précision

- **Le Rappel (Recall) :** Le rappel mesure *la capacité du système à retrouver tous les documents pertinents*.

Autrement dit :

« Parmi tous les documents pertinents qui existent dans la collection, combien le système en a-t-il retrouvés ? »

- **La Précision** La précision mesure *la capacité du système à ne sélectionner que des documents pertinents*.

En d'autres termes :

« Parmi tous les documents que le système a retournés, combien sont vraiment pertinents ? »

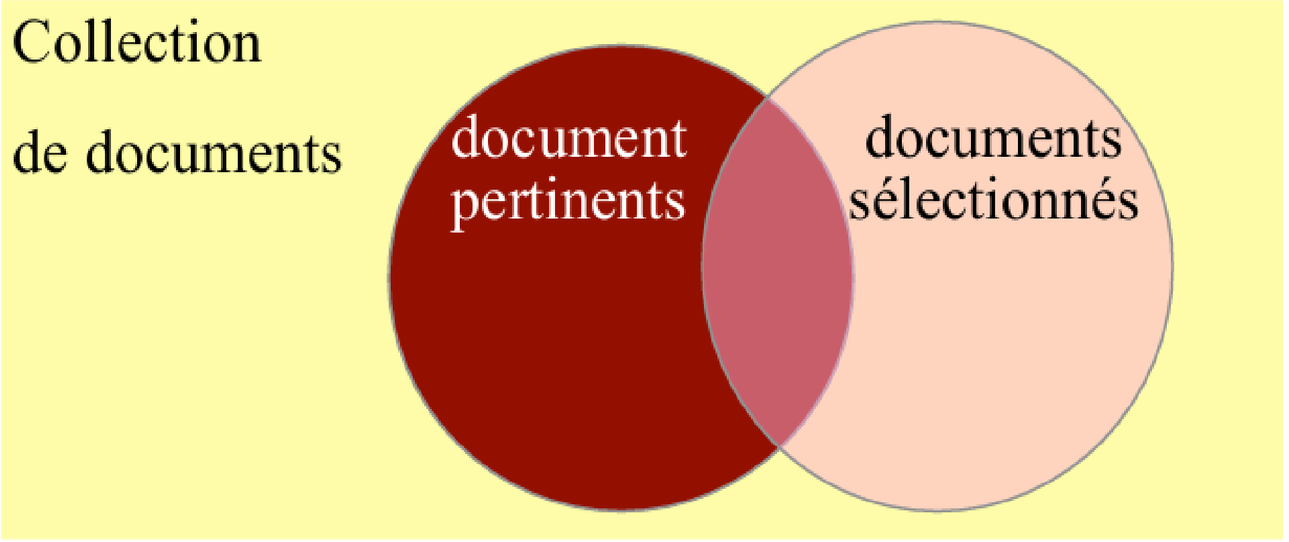
IV. Critères d'évaluation

IV.1 le rappel et la précision

$$\text{Pre}cision = \frac{Nombre\ de\ documents\ pertinents\ selectionnes}{Nombre\ total\ de\ documents\ selectionnes}$$

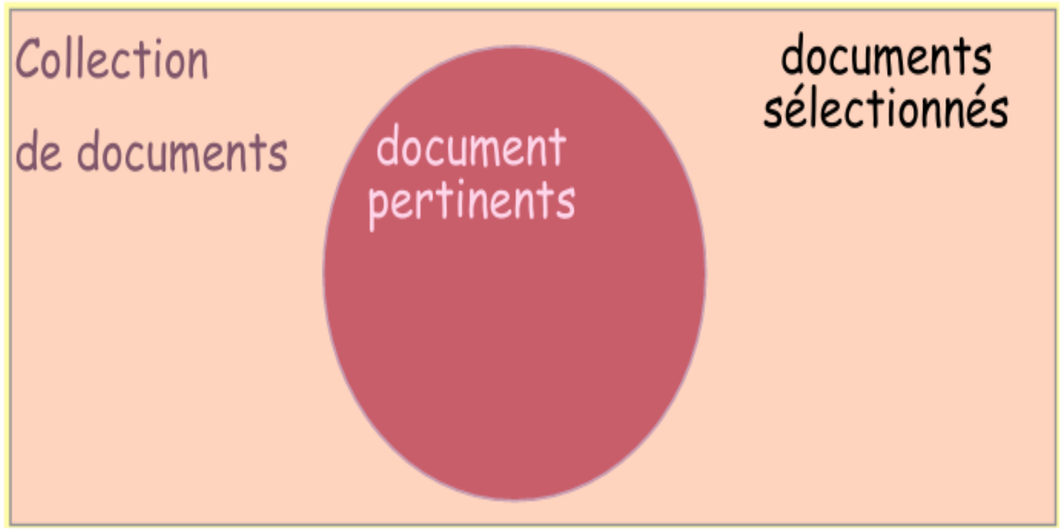
$$\text{Rappel} = \frac{Nombre\ de\ documents\ pertinents\ selectionnes}{Nombre\ total\ de\ documents\ pertinents}$$

	Sélectionnés	**Non sélectionnés**
Pertinents	**Sélection & Pertinent**	Non sélection mais Pertinent
Non pertinents	Sélection & Non pertinent	Non sélection & Non pertinent



IV. Critères d'évaluation

IV.1 le rappel et la précision : Pourquoi deux facteurs



- Il est **très facile d'obtenir un rappel élevé** : Il suffit de sélectionner *toute la collection*. Mais la **précision devient très faible**.

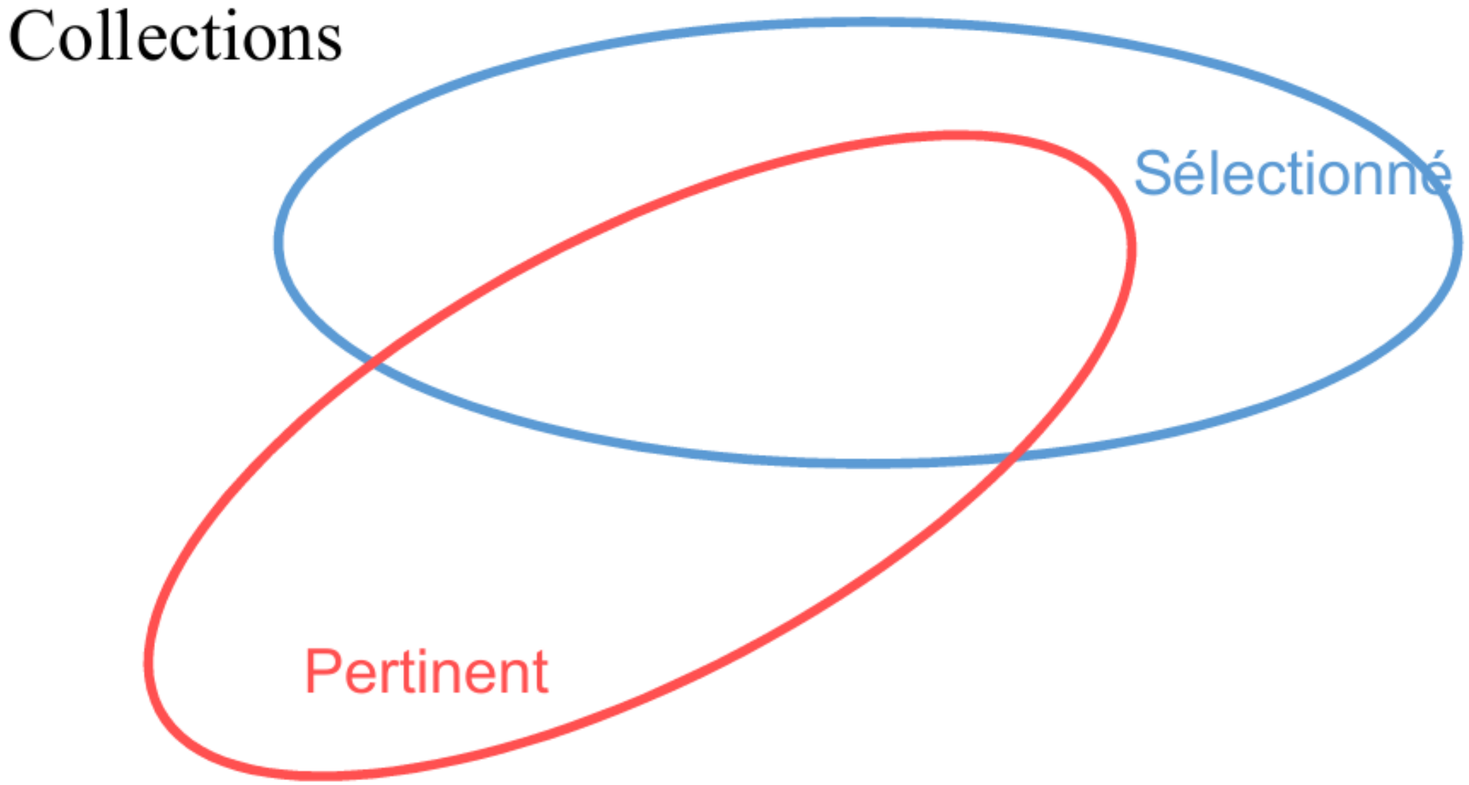
- À l'inverse, on peut avoir une **précision très élevée**: En sélectionnant seulement quelques documents très sûrs. Mais le **rappel devient très faible**.

- C'est pourquoi on utilise les deux mesures pour évaluer correctement un SRI. elles sont complémentaires.

IV. Critères d'évaluation

IV.1 le rappel et la précision : Pertinent vs. Sélectionné

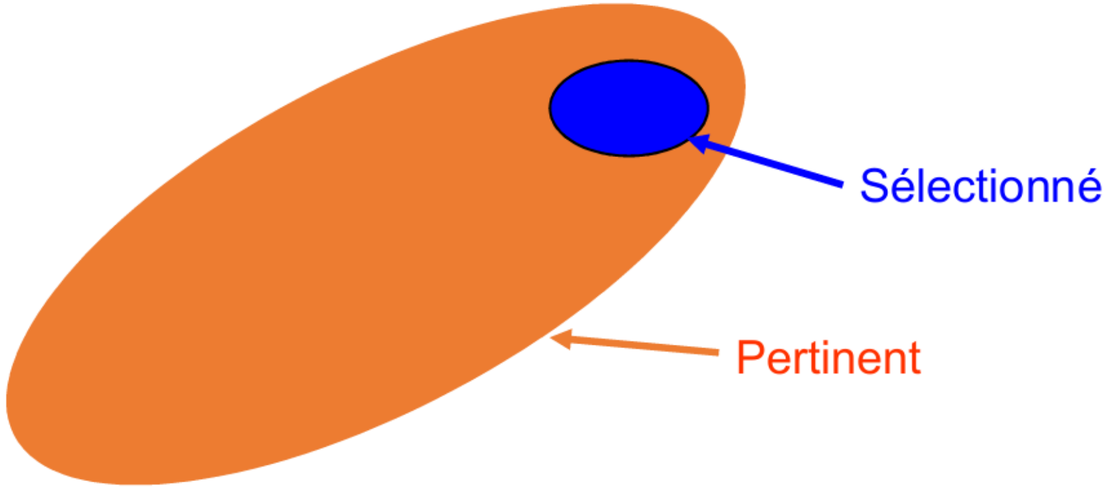
Pour bien visualiser cette complémentarité, regardons plusieurs scénarios concrets.



IV. Critères d'évaluation

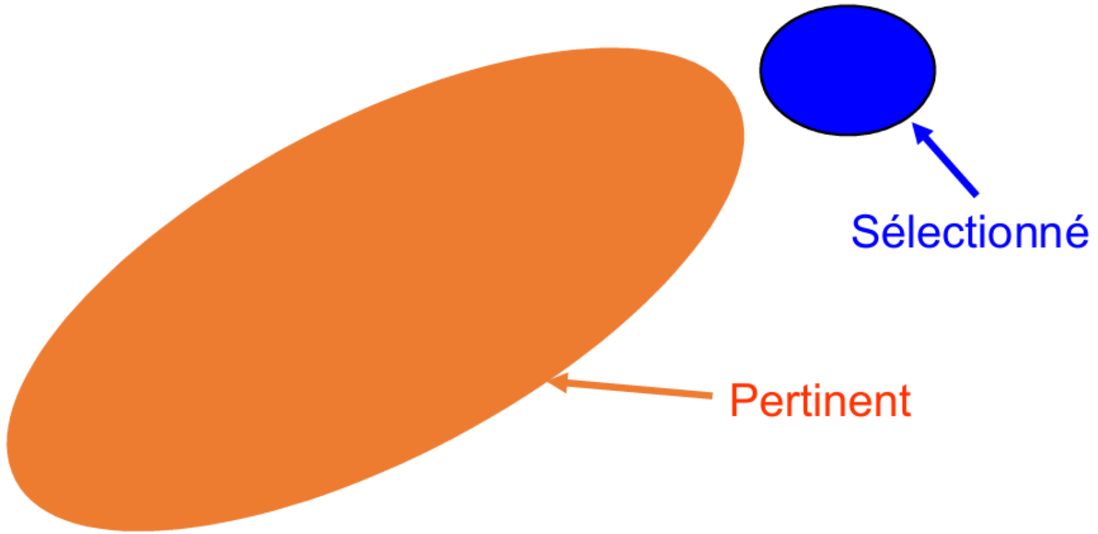
IV.1 le rappel et la précision : Pertinent vs. Sélectionné

Précision très élevée, rappel très faible



Il retrouve très peu de documents pertinents → rappel faible.
Et parmi ce qu'il sélectionne, presque tout est pertinent → précision élevée.

Précision très faible, rappel très faible null

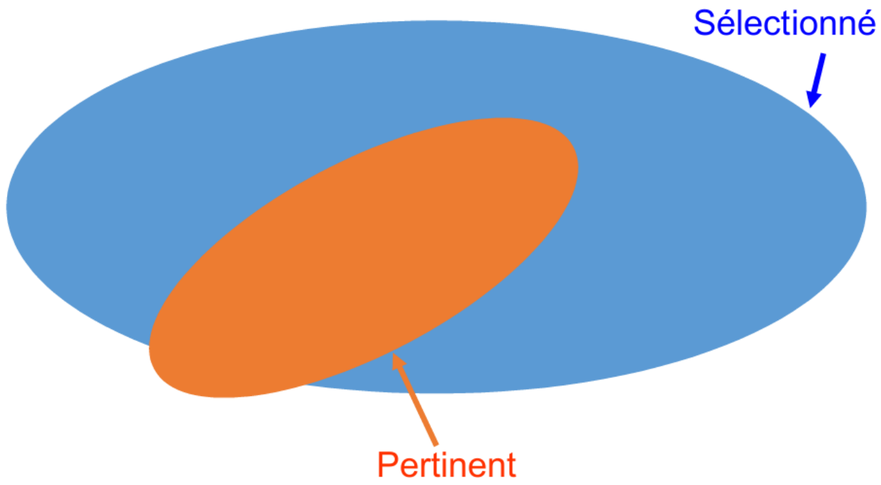


Il retrouve très peu de documents pertinents → rappel faible.
Et parmi ce qu'il sélectionne, presque rien n'est pertinent → précision faible.

IV. Critères d'évaluation

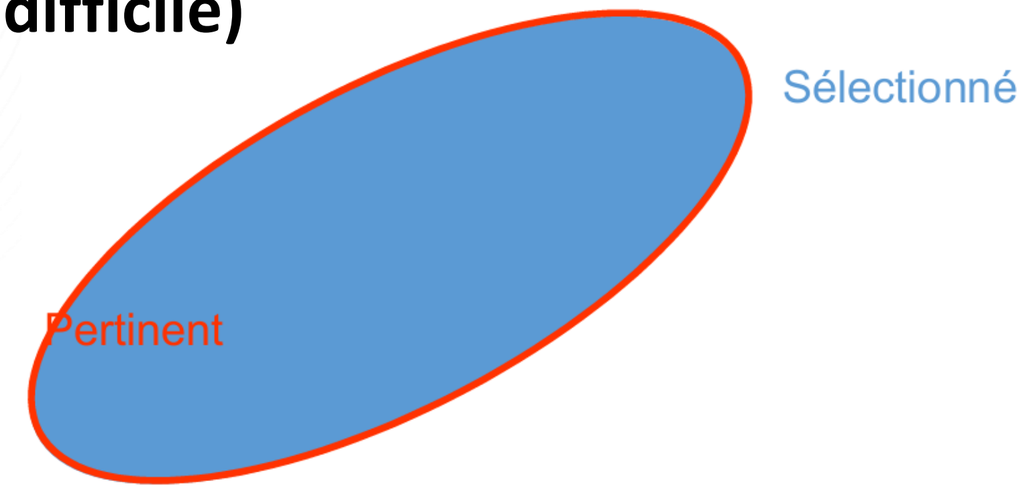
IV.1 le rappel et la précision : Pertinent vs. Sélectionné

Rappel élevé, mais précision faible



Il retrouve la plupart des documents pertinents → rappel élevé.
Mais il renvoie aussi beaucoup de documents non pertinents → précision faible.

Précision élevée, rappel élevé (idéal, mais difficile)



Il retrouve la plupart des documents pertinents → rappel élevé, et parmi eux beaucoup sont pertinents → précision élevée.

IV. Critères d'évaluation

IV.1 le rappel et la précision : Exemple

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**

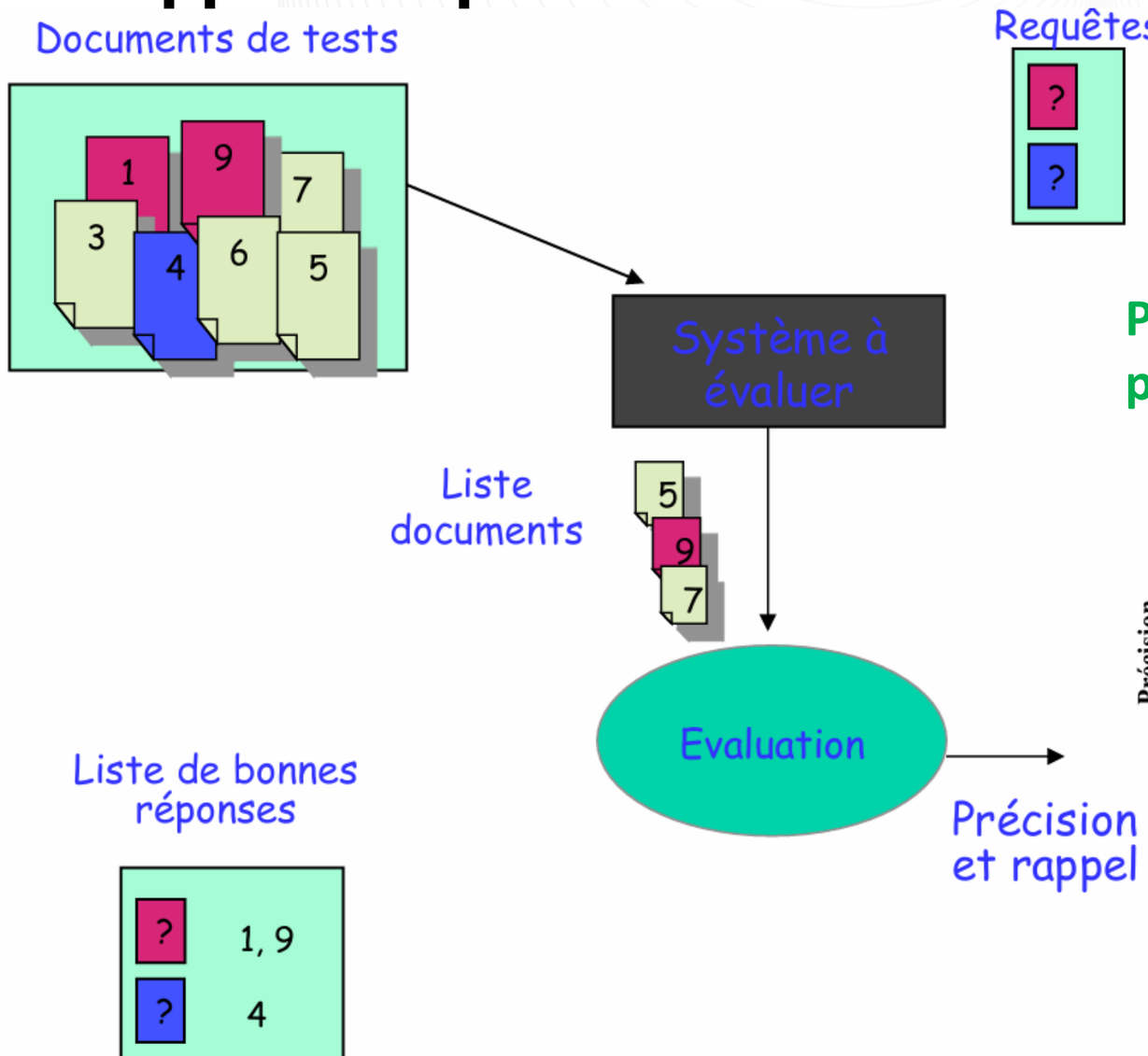
Calculer la précision et le rappel pour les deux systèmes A et B.
Le quel des deux systèmes est meilleur ?

Solution : Rappel final A =2/4 = 0.5
Précision final A = 2/4 = 0.5
Rappel final B =3/4 = 0.75
Précision final B = 3/7 = 0.428

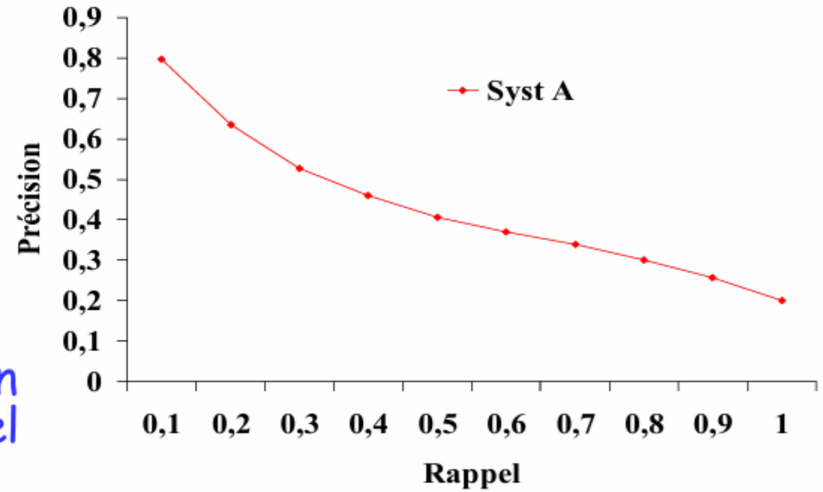
On remarque que :
Du point de vue rappel le système B qui est meilleur.
Du point de vue précision c'est le système A qui meilleur.

IV. Critères d'évaluation

IV.1 le rappel et la précision : Processus de calcul



Pour chaque document, le rappel et la précision sont calcules progressivement



IV. Critères d'évaluation

IV.1 le rappel et la précision : Exercice

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.

IV. Critères d'évaluation

IV.1 le rappel et la précision : Exercice

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

On utilise ces calculs progressifs pour tracer la courbe rappel-précision

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

IV. Critères d'évaluation

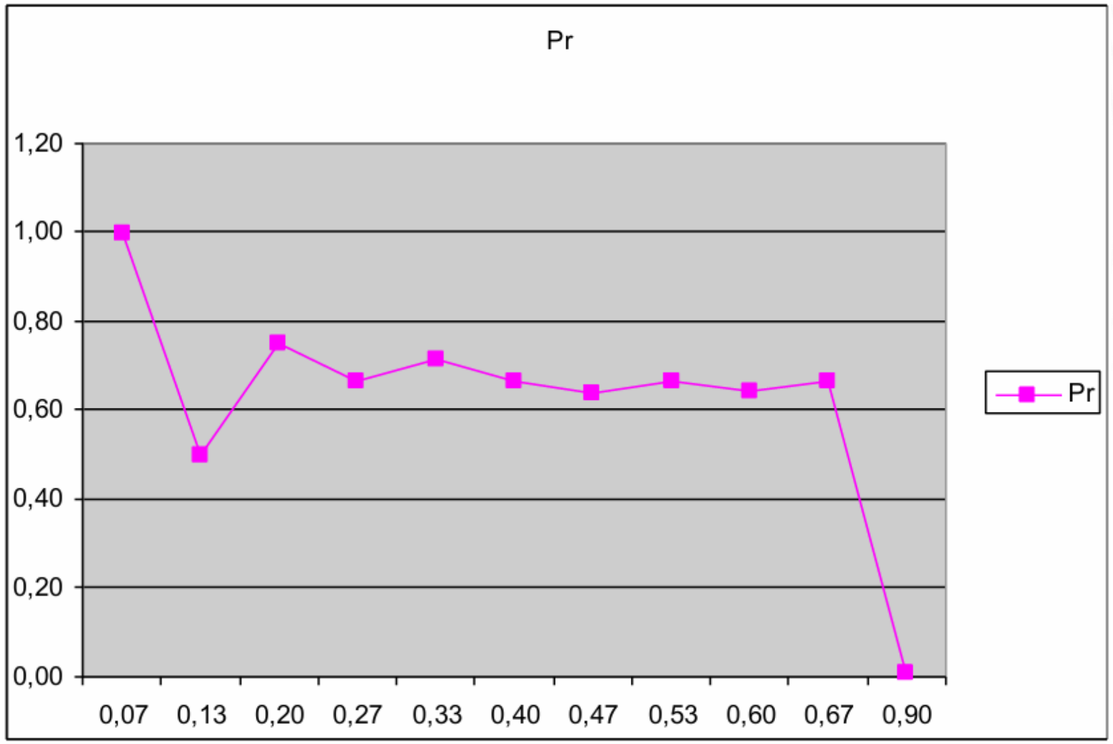
IV.2 Courbe Rappel / Précision pour une requête

- Une représentation graphique qui montre la relation entre **rappel** et **précision**. Chaque point correspond à un document pertinent retrouvé par le système.
- Permet de **visualiser le compromis rappel / précision**
- **Courbe haute et plate** : le système retrouve rapidement la plupart des documents pertinents → très performant.
- **Courbe qui chute rapidement** : le système sélectionne beaucoup de documents non pertinents avant de retrouver les pertinents → performance moindre.

IV. Critères d'évaluation

IV.2 Courbe Rappel / Précision pour une requête : Exemple

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

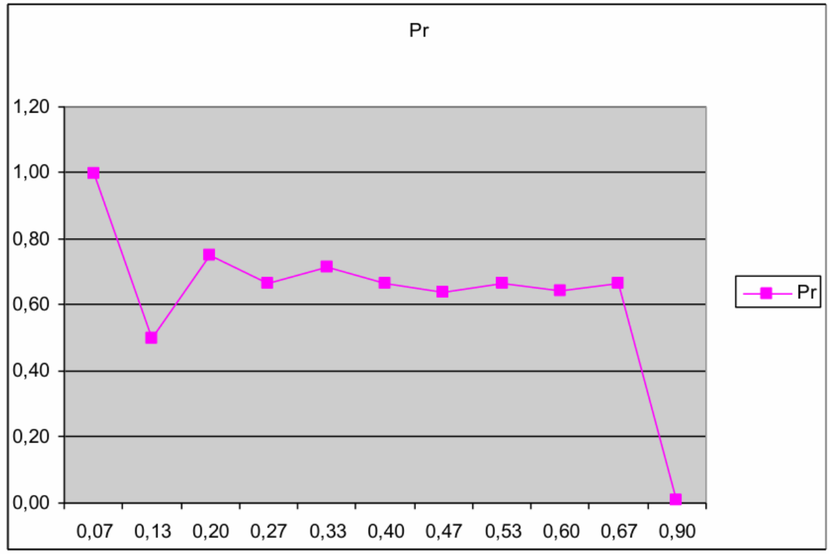


- Les points représentent les valeurs mesurées à chaque document pertinent.
- La courbe montre le compromis **précision vs rappel**.
- Permet d'évaluer globalement la performance d'un SRI pour cette requête.

IV. Critères d'évaluation

IV.3 Interpolation de la Courbe Rappel / Précision

- On peut observer que les points ne forment pas forcément une courbe régulière : la précision monte, descend, …. irrégulière (montées/baisses).
- Ce comportement rend la comparaison entre plusieurs systèmes un peu difficile,
- Dans le but d'avoir une représentation plus stable, plus lisible et surtout **plus comparable**, on applique ce qu'on appelle ***l'interpolation*** de la courbe Rappel/Précision.



IV. Critères d'évaluation

IV.3 Interpolation de la Courbe Rappel / Précision

- Idee : Pour chaque niveau de rappel standard, on ne garde pas la précision exacte mesurée, mais **la précision maximale atteinte pour tous les rappels supérieurs ou égaux à ce niveau**.
- On calcule une précision interpolée pour chaque point de rappel :
  r ∈ {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 rappel rj est la valeur maximale des précisions observées pour tous les rappels r tels que r ≥ rj.

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

- **Cela permet d'obtenir une courbe plus lisse, plus monotone, et surtout beaucoup plus facile à comparer entre requêtes ou entre systèmes.**

IV. Critères d'évaluation

IV.3 Interpolation de la Courbe Rappel / Précision : Exemple

Rappel	Précision
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

On veut **interpôler la précision** pour chaque rappel standard : 0.0, 0.1, 0.2, … , 1.0

La règle est :

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

Donc pour chaque rappel rj, on regarde toutes les précisions des rappels **r ≥ rj**, et on prend **la plus grande**.

Rappel standard	Précision interpolée
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

IV. Critères d'évaluation

IV.4 Précision moyenne pour une requête

- Lorsqu'on évalue un SRI, on souhaite souvent **une valeur unique** représentant la performance pour une requête.
- Cette valeur permet de Comparer différents systèmes.

1. Précision moyenne non interpolée (PrecAvg)

- Elle combine **précision et rappel** sur tous les documents pertinents.
- Chaque fois qu'un document pertinent est retrouvé, on calcule la précision jusqu'à ce rang.
- On fait ensuite la **moyenne de toutes ces précisions** sur le nombre de documents pertinents donnés par l'environnement de tests.

IV. Critères d'évaluation

IV.4 Précision moyenne pour une requête

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	x

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

R=6/6 , P=6/14=0.42

Calcul de la précision moyenne :

AP=AvgPrec

$$= \frac{1 + 1 + 0{,}75 + 0{,}667 + 0{,}38 + 0.42}{6} \approx 0{,}702$$

Même chose pour la Précision moyenne interpolée

IV. Critères d'évaluation

IV.5 Autres mesures de moyennes

F-Mesure

- La **F-Mesure** combine **précision (P)** et **rappel (R)** en une seule valeur.
- Elle est définie comme **la moyenne harmonique** de P et R
- Introduite par **van Rijsbergen, 1979**.
- Évite que de fortes différences entre P et R soient masquées par une simple moyenne arithmétique.
- Idéal pour résumer la performance globale d'un système sur une requête.

$$F = \frac{2 \cdot P \cdot R}{P + R}$$

IV. Critères d'évaluation

IV.5 Autres mesures de moyennes

E-Mesure (F-Mesure paramétrique)

- L'**E-Mesure** est une variante **paramétrée** de la F-Mesure.
- Elle permet de **modifier le poids de la précision par rapport au rappel** grâce au paramètre β:

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

Rôle de β:

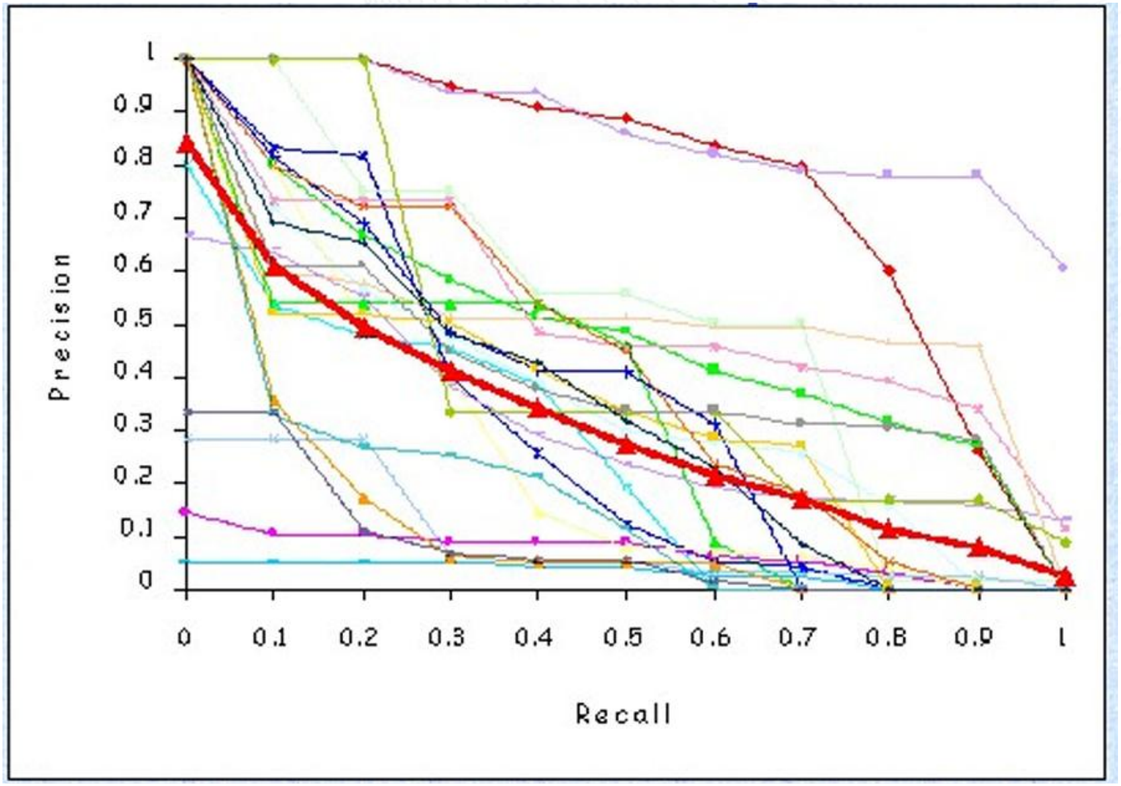
$\beta = 1$: poids égal à la précision et au rappel $\rightarrow E = F$

$\beta > 1$: privilégie le rappel

$\beta < 1$: privilégie la precision

IV. Critères d'évaluation

IV.6 R-P courbes sur l'ensemble des requêtes



- Les courbes rappel-précision pour **chaque requête** sont difficiles à comparer.
- Les résultats deviennent illisibles lorsque l'on a **plusieurs requêtes ou plusieurs systèmes**.
- Pour disposer d'une mesure synthétique et comparable, on calcule alors **une moyenne des performances obtenues sur l'ensemble des requêtes**.

- La mesure la plus utilisée pour cela est la **Mean Average Precision (MAP)**, qui représente la **moyenne des précisions moyennes (AP)** calculées pour chaque requête.

IV. Critères d'évaluation

IV.7 Précision Moyenne – Mean Average Precision – MAP

- La **précision moyenne (AP)** pour une requête q, ayant des documents pertinents $\{d_1, \ldots, d_m\}$,est calculée en faisant la moyenne des précisions obtenues aux positions où apparaissent les documents pertinents dans la liste de résultats.
- La **précision moyenne globale (MAP)** est la moyenne des AP calculées sur l'ensemble des requêtes Q utilisées pour évaluer un système de recherche d'information.
- La MAP s'applique aux **valeurs précision–rappel non interpolées**.

IV. Critères d'évaluation

IV.7 Précision Moyenne - Mean Average Precision - MAP

Exemple

Requête 1

Ra	Pr
0.125	1
0.25	0.5
0.375	0.5
0.5	0.444
0.625	0.455
0.75	0.429

Requête 2

Ra	Pr
0.125	0.5
0.25	0.5
0.375	0.429
0.5	0.444
0.625	0.500
0.75	0.462

AP Average precision for each query

$$\text{AvgPrec}_1 = \frac{1 + 0.5 + 0.5 + 0.444 + 0.455 + 0.429}{6}$$

AvgPrec1=3.328/6≈0.555

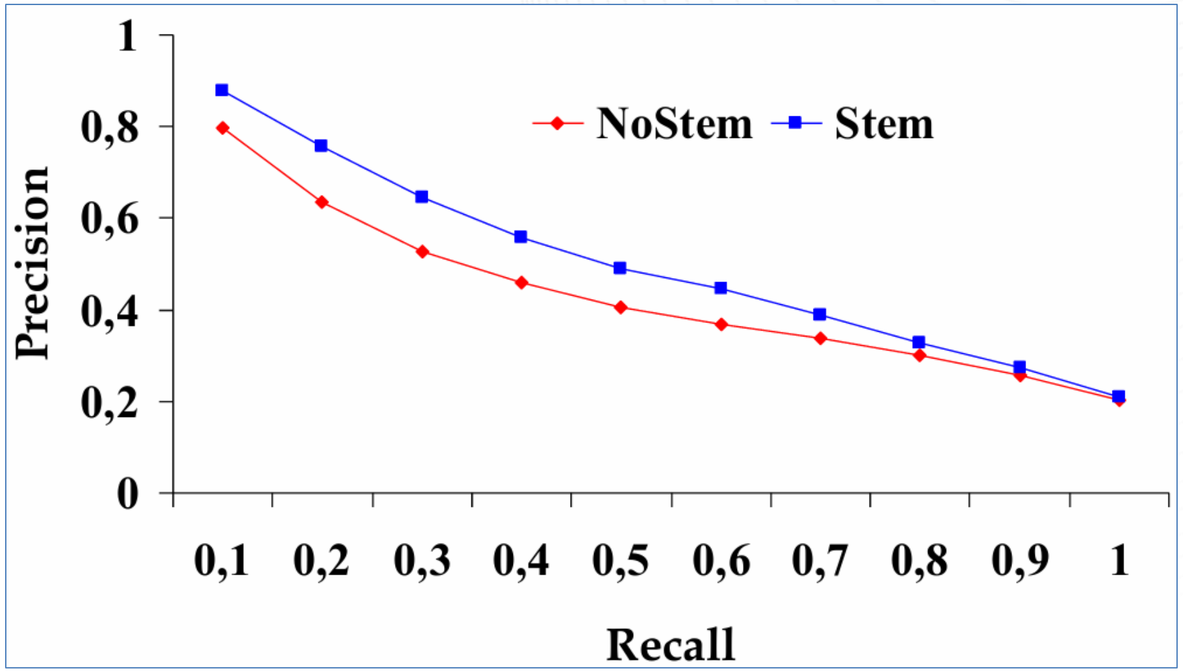
AvgPrec_2 = 2.835/6≈0.4725

MAP for all queries

$$MAP = \frac{0.555 + 0.4725}{2} \approx 0.5138$$

IV. Critères d'évaluation

IV.8 Comparaison de deux systèmes sur un ensemble de requêtes



- La comparaison visuelle entre les SRI se fait généralement à l'aide des courbes précision–rappel interpolées, moyennées sur l'ensemble des requêtes.
- Ces courbes sont moyennées sur l'ensemble des requêtes pour obtenir une vue globale.
- **Courbe plus haute** = meilleure performance.

- Ce graphique permet de conclure que, **sur l'ensemble des requêtes testées**, le système avec stemming est supérieur en précision à rappel égal.
- On peut calculer **(MAP)** pour avoir une mesure numérique unique de la supériorité de Stem sur NoStem.

V. 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 les mesures les plus utilisées :

1. **Precision au Rang X (Precision at rank X) P@X**
2. **R-Précision (R-Precision)**
3. **Rang réciproque (Reciprocal Rank) RR**
4. **Gain Cumulé (Discounted Cumulative Gain) DCG**
5. **Gain Cumulé Normalisé (Normalized Discounted Cumulative Gain) nDCG**

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

V. I. Precision au Rang X (P@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 15 docs = 6/15

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

V. 2. R-Précision (R-Precision)

- La R-Précision est donc la précision mesurée après avoir examiné les **R premiers documents retournés par le système.**
- La **R-précision** correspond à **P@k**, où k est égal au **nombre total de documents pertinents** dans l'environnemenpour la requête t de test.

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

In this sample the R-precision = P@8

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

V. 3. Rang réciproque (Reciprocal Rank)

Le **Reciprocal Rank (RR)** est une mesure qui évalue **à quelle position apparaît le premier document pertinent** dans la liste des résultats.

Plus le **premier document pertinent** apparaît tôt, meilleure est la valeur de RR.

$$RR = \frac{1}{rang\ du\ premier\ document\ pertinent}$$

Cette mesure est particulièrement utilisée pour :

- **Question Answering (QA)**
- **Navigational search**
- **Systèmes où un seul document (ou peu) ont vraiment de la valeur**

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

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

V. 4. Le Gain Cumulé (Discounted Cumulative Gain) DCG

Evalue dans quelle mesure un système place les documents pertinents en haut de la liste des résultats. Elle prend en compte deux aspects :

La pertinence du document : plus un document est pertinent, plus il contribue au gain total.

La position du document dans la liste : un document pertinent situé en bas de la liste apporte beaucoup moins de gain qu'un document situé en haut.

Formule donnée par :

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

- rel_i : RSV ou score de pertinence du document à la position i (donné par l'environnement de test)
- p : profondeur d'évaluation (nombre de documents considérés dans la liste)
- log2 (i) = ln(i) / ln(2)

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

V. 4. Le Gain Cumulé (Discounted Cumulative Gain) DCG

Soit une liste de documents retournés par un système A :
d1, d2, d3, d4, d5, d6, d7, d8, d9, d10

Degrés de pertinence (rsv) selon l'environnement de tests :

- rsv(d1) = 0.3
- rsv(d2) = 0.2
- rsv(d3) = 0.3
- rsv(d6) = 0.4
- rsv(d7) = 0.5
- rsv(d9) = 0.3
- Tous les autres = 0

Questions :

1- Calculez la DCG_{10}

2- Quelle est la meilleure valeur de DCG que peut retourner ce système dans le meilleur cas

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

V. 4. Le Gain Cumulé (Discounted Cumulative Gain) DCG

Solution

1- Calcul de DCG_{10}

On calcule document par document :

$$d_1 = 0.3$$
$$d_2 = 0.2/\log_2(2) = 0.2/1 = 0.2$$
$$d_3 = 0.3/\log_2(3) \approx 0.3/1.58496 \approx 0.1894$$
$$d_4 = 0/\log_2(4) = 0$$
$$d_5 = 0/\log_2(5) = 0$$
$$d_6 = 0.4/\log_2(6) \approx 0.4/2.58496 \approx 0.1548$$
$$d_7 = 0.5/\log_2(7) \approx 0.5/2.8074 \approx 0.1781$$
$$d_8 = 0/\log_2(8) = 0$$
$$d_9 = 0.3/\log_2(9) \approx 0.3/3.1699 \approx 0.0946$$
$$d_{10} = 0/\log_2(10) = 0$$

$$\boldsymbol{DCG_{10} \approx 0.3 + 0.2 + 0.1894 + 0 + 0 + 0.1548 + 0.1781 + 0 + 0.0946 + 0}$$
$$\boldsymbol{DCG_{10} \approx 1.1169\ (\approx 1.12)}$$

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

V. 4. Le Gain Cumulé (Discounted Cumulative Gain) DCG

Solution

2- Quelle est la meilleure valeur de DCG que peut retourner ce système dans le meilleur cas : Pour obtenir le **DCG maximal**, il faut placer les documents dans l'ordre **décroissant de pertinence** :

Ordre optimal : $d_7(0.5), d_6(0.4), d_1(0.3), d_3(0.3), d_9(0.3), d_2(0.2), d_4(0), d_5(0), d_8(0), d_{10}(0)$

Calcul :

$$d_7 = 0.5$$

$$d_6 = 0.4/\log_2(2) = 0.4$$

$$d_1 = 0.3/\log_2(3) \approx 0.1894$$

$$d_3 = 0.3/\log_2(4) = 0.3/2 = 0.15$$

$$d_9 = 0.3/\log_2(5) \approx 0.3/2.3219 \approx 0.1292$$

$$d_2 = \frac{0.2}{\log_2(6)} \approx \frac{0.2}{2.58496} \approx 0.0774$$

$$\boldsymbol{DCG_{10}^{max} \approx 0.5 + 0.4 + 0.1894 + 0.15 + 0.1292 + 0.0774 \approx 1.445}$$

V. Mesures focalisées sur le “top” de la liste

V. 4. Le Gain Cumulé (Discounted Cumulative Gain) DCG

Un **DCG élevé** signifie :

- Les documents pertinents sont plutôt en haut de la liste.
- Les documents fortement pertinents (ex. score 0.5) ne sont pas trop pénalisés.

Un **DCG faible** signifie :

- Les documents pertinents sont dispersés plus bas dans la liste.
- Leur gain est “discounted” par la position.

V. Mesures focalisées sur le “top” de la liste

V. 5. DCG Normalizé (NDCG)

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

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

- **DCGp** : score du système
- **IDCGp** : meilleur DCG possible (Reclassement idéal en se basant sur les Valeur RSV),
- Le nDCG est **toujours entre 0 et 1**.

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

V. 5. DCG Normalizé (NDCG)

Les documents triés du plus pertinent au moins pertinent, selon les degrés de pertinence de l'environnement de tests.

Exemple:

DCG10=1.117

Doc	rsv
d7	0.5
d6	0.4
d1	0.3
d3	0.3
d9	0.3
d2	0.2
les autres	0

Calcul de $IDCG_{10}$ (DCG idéal)

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

$$IDCG_{10} = 1.735$$

Calcul du nDCG

$$nDCG_{10} = \frac{1.117}{1.735} = 0.644$$

VI. Comparaison des systèmes

Pour comparer deux systèmes **A** et **B**, on peut utiliser leurs valeurs de mesures d'évaluation :

$$\text{Gain}(\%) = \frac{Val(A) - Val(B)}{Val(B)} \times 100$$

Si l'amélioration est **≥ 5%**, on considère généralement que **A est réellement meilleur que B**.

Comparer leurs courbes

On compare les courbes **Précision-Rappel,** Si la courbe du système **A** est **toujours au-dessus** de celle de **B**, alors A est meilleur sur cette collection.

VI. Comparaison des systèmes

Que se passe-t-il quand on change de collection ?

Quand on change de collection :

1. Les performances peuvent changer

Les résultats (MAP, NDCG, Précision, etc.) ne seront plus forcément les mêmes.

2. L'ordre des systèmes peut s'inverser

Un système **A** peut être meilleur dans une collection… mais **moins bon** sur une autre.

Pourquoi ?

Parce que les performances dépendent :

- du type de documents (longs, courts, spécialisés, généraux)
- du domaine (médical, actualités, web…)
- du vocabulaire
- de la difficulté des requêtes ..etc

3. Aucun système n'est universellement meilleur

VII. Avantages et inconvénients des collections de tests

Avantages

- Elles permettent d'obtenir des mesures de performance fiables.
- Facilite la comparaison des systèmes 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