UNIVERSITÉ DES SCIENCES ET TECHNOLOGIE HOUARI BOUMEDIENE

FACULTÉ D'INFORMATIQUE - MASTER 2 SII

RECHERCHE D'INFORMATION
INFORMATION RETRIEVAL

CHAPITRE 6: Modèles Probabilistes BIR , BIR pondéré, BM25

03 Novembre 2025

I. INTRODUCTION

Pourquoi utiliser les probabilités en RI ?

La recherche d'information (RI) est un processus incertain :

- Les requêtes d'utilisateurs sont souvent imprécises ou ambiguës.
- Les documents peuvent être partiellement pertinents.
- Il existe une incertitude dans la représentation du contenu et des besoins.

La théorie des probabilités permet de mesurer cette incertitude et imprécision et de modéliser le "degré de pertinence".

I. INTRODUCTION

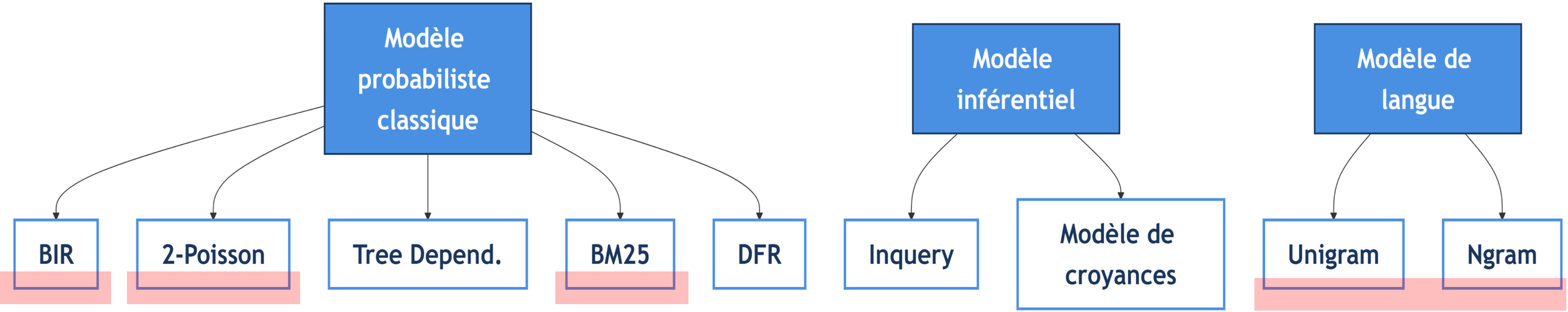
Le modèle probabiliste tente d'estimer la probabilité d'observer des événements liés au document et à la requête.

Plusieurs modèles probabilistes, se différencient selon

1) Les événements qu'ils considèrent
   - P(pertinence/d, q) : probabilité de pertinence de d vis à vis de q
   - P(q,d) : probabilité d'observer simultanément la requête et le document.
   - P(q|d)
   - P(d|q)
   - ……
2) Les distributions (lois) qu'ils utilisent.

I. INTRODUCTION

LA RECHERCHE D'INFORMATION BASE SUR LES PROBABILITÉS



II. Rappels de probabilités : Notions

- P(A) = probabilité que A se produise
- $P(non\ A) = 1 - P(A).$

  Exemple : P("pile") = P("face") = ½
- Somme des probabilités : Σ P(s) = 1
- P(A∪B) = P(A) + P(B) – P(A∩B)
- Probabilité conditionnelle : P(A|B)

Exemple : P(“retrieval” | “information”) > P(“retrieval” | “politic”)

“retrieval” est plus probable dans un texte contenant “information” que dans un texte parlant de “politic”.

II. Rappels de probabilités : Distribution

Une *distribution de probabilités* **décrit la répartition des probabilités sur tous les événements possibles d'une expérience.**

Exemple:

- P(RED) = probabilité de tirer une boule rouge
- P(BLUE) = probabilité de tirer une boule bleue
- P(ORANGE) =…etc.

Dans nos modèles de recherche d'information, c'est pareil : on cherche la distribution des probabilités de pertinence pour chaque document ou mot.

II. Rappels de probabilités : Estimation de Distribution

La probabilité est estimée à partir des observations : combien de fois chaque événement s'est produit.

$$P(evenement) = \frac{nombre\ de\ cas\ favorables}{nombre\ total\ de\ cas}$$

Deux conditions :

- $0 \leq P \leq 1$
- $\sum_i P(s_i) = 1$

Exemple :

Sur 100 tirages :

- 40 RED→ P(RED)=0.4
- 35 BLUE → P(BLUE)=0.35
- 25 ORANGE → P(ORANGE)=0.25

- En RI, on estime P(mot|document) à partir de la fréquence du mot dans le document.
- Plus un mot apparaît souvent, plus P(mot|document) est élevée.

II. Rappels de probabilités : Probabilité conditionnelle et Règle de Bayes

- **Événements indépendants :**

$$\boldsymbol{P}(\mathrm{A}, \mathrm{B}) = \boldsymbol{P}(\boldsymbol{A}) \times \boldsymbol{P}(\boldsymbol{B})$$

Ex: lancer deux des

- **Événements dépendants :**

$$P(\mathrm{A}, \mathrm{B}) = P(A) \times P(B \mid A)$$

$$P(\mathrm{A}, \mathrm{B}, \mathrm{C}) = P(A) \times P(B \mid A) \times P(C \mid \mathrm{A}, \mathrm{B})$$

Ex: tirer plusieurs boules sans remise.

- **Règle de Bayes :**

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

II. Rappels de probabilités : Variable aléatoire

Une **variable aléatoire (V.A.)** est une fonction qui associe à chaque résultat d'une expérience aléatoire un **nombre réel**.

$$X: \Omega \to \mathbb{R}, \omega \mapsto X(\omega)$$

- Ω: ensemble des résultats possibles (l'univers).
- $X(\omega)$: valeur numérique associée au résultat ω.

Exemple : Expérience : lancer deux dés (un bleu et un rouge)

$$\Omega = \{(b = 1, r = 1), (b = 1, r = 2), \dots, (b = 6, r = 6)\}$$

On définit la variable aléatoire S= somme des deux dés.

$$S(\omega) = b + r$$

Les valeurs possibles : $S \in \{2, 3, \dots, 12\}$

Distribution de probabilité

$$P(S = 2) = \frac{1}{36}, P(S = 3) = \frac{2}{36}, \dots, P(S = 7) = \frac{6}{36}$$

II. Rappels de probabilités : Variable aléatoire

Types de variables aléatoires

- **Discrète :** valeurs dénombrables (ex. lancer de dés).
- **Continue :** valeurs dans un intervalle (ex. temps d'attente, taille).

Loi de probabilité d'une variable aléatoire (discrète)

Décrit la probabilité associée à chaque valeur possible de cette variable.

$$p_i = P(X = x_i), \qquad avec\ 0 \leq p_i \leq 1\ et \sum_i p_i = 1$$

II. Rappels de probabilités : Variable aléatoire

Loi uniforme

Une variable aléatoire X suit une loi uniforme discrète si toutes ses valeurs possibles ont la même probabilité.

$$P(X = k) = \frac{1}{n} \quad pour \quad X \in \{1, 2, \dots, n\}$$

Exemple : Soit une variable aléatoire : X= résultat du lancer du dé , Ensemble des valeurs possibles : $X \in \{1, 2, 3, 4, 5, 6\}$, Il y a **n = 6** issues possibles, toutes **équiprobables**

Donc : $P(X = k) = \frac{1}{6}$

Pour chaque valeur :

$$P(X = 1) = P(X = 2) = \dots = P(X = 6) = \frac{1}{6}$$

II. Rappels de probabilités : Variable aléatoire

Loi de Bernoulli

Une variable aléatoire X suit une **loi de Bernoulli** si elle ne peut prendre que deux valeurs X={0,1} :

$$X = \begin{cases} 1 & avec\ probabilité\ p \\ 0 & avec\ probabilité\ 1-p \end{cases}$$

$$P(X = x) = p^x(1-p)^{1-x}\ x \in \{0, 1\},$$

Exemple :

Lancer une pièce de monnaie :

- $X = 1$ si on obtient “pile” P(X=1)=0,5
- $X = 0$ si on obtient “face” P(X=0)=1−0,5=0,5

II. Rappels de probabilités : Variable aléatoire

Loi binomiale

Une variable aléatoire X suit une **loi binomiale** si elle compte le **nombre de succès** obtenus lors de n répétitions indépendantes d'une expérience de Bernoulli de probabilité p.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Avec le Coefficient binomial :

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Exemple :

On lance une pièce n=10 fois et on compte le nombre de "pile". Si $p = 0{,}5$, alors la probabilité d'obtenir exactement k=6 piles est :

$$P(X = 6) = \binom{10}{6} (0.5)^6 (0.5)^4 = 210 \times (0.5)^{10} \approx 0.205$$

II. Rappels de probabilités : Variable aléatoire

Loi multinomiale

C’est une généralisation de la loi binomiale à plusieurs issues possibles (m au lieu de 2).

$$P(X_1 = k_1, X_2 = k_2, \ldots, X_m = k_m) = \frac{n!}{k_1! k_2! \ldots k_m!} p_1^{k_1} p_2^{k_2} \ldots p_m^{k_m}$$

avec $\sum k_i = n$ et $\sum p_i = 1$.

II. Rappels de probabilités : Variable aléatoire

Loi de Poisson

Une variable aléatoire X suit une **loi de Poisson** si elle modélise le **nombre d'occurrences** d'un événement rare dans un intervalle donné.

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- λ: moyenne (nombre attendu d'occurrences)
- k: nombre observé

Exemple :

Une banque reçoit en moyenne $\lambda = 5$ appels par minute.

Quelle est la probabilité d'en recevoir exactement k=7 ?

$$P(X = 7) = \frac{5^7 e^{-5}}{7!} \approx 0.104$$

III. Probability Ranking Principle PRP : Du calcul de probabilités au modèle probabiliste

- ✓ Les modèles probabilistes s'appuient sur les lois de probabilité pour estimer la *pertinence* des documents pour une requête donnée.
- ✓ L'objectif est d'estimer la probabilité qu'un document d soit **pertinent (Relevent)** pour une requête q. $\boldsymbol{P(R \mid \mathbf{d}, \mathbf{q})}$

Ou **R**: l'événement « le document est pertinent »

d: un document de la collection

q: la requête utilisateur.

III. Probability Ranking Principle PRP : Du calcul de probabilités au modèle probabiliste

Le modèle probabiliste considère deux **classes de documents** :

- **Classe des documents pertinents (R)** exprimé par : **P(R|d,q)**

  Documents qui répondent réellement au besoin exprimé par la requête.

- Classe des documents non pertinents (NR): **P(NR|d,q)**

  Documents qui contiennent éventuellement les mots de la requête, mais ne répondent pas au besoin d'information.

« Que faire de ces probabilités ? »

III. Probability Ranking Principle PRP : Du calcul de probabilités au modèle probabiliste

Pourquoi mesurer la Pertinence et la Non pertinence

- Parce qu'il existe toujours **une incertitude** : un document peut ressembler à un texte pertinent sans réellement répondre à la requête.
- En modélisant **les deux classes (R et NR)**, on peut estimer **le degré de confiance** dans la pertinence de chaque document.
- Cela permet d'établir un **classement probabiliste** plus robuste qu'un simple comptage de mots communs.

Autrement dit : on ne cherche pas à dire "ce document est pertinent ou non", mais plutôt "ce document a 80 % de chances d'être pertinent, et 20 % de ne pas l'être".

III. Probability Ranking Principle PRP

> ‘’If a retrieval system’s response to each request is a **ranking of the documents** in **order of decreasing probability of relevance**, where the probabilities are estimated as **accurately** as possible from all **available evidence**, then the **overall effectiveness** of the system will be the **best that is obtainable**.’’
>
> Stephen E. Robertson, 1977

- La pertinence est incertaine, donc on la modélise par une probabilité.
- Les documents sont triés par ordre décroissant de P
- Si ces probabilités sont bien estimées, alors le classement obtenu est le plus performant possible.
- La PRP constitue la fondation théorique de tous les modèles probabilistes modernes (BIR, BM25, modèles de langue, etc.).

Les documents ne sont pas classés selon une seule mesure fixe, mais selon leur probabilité globale d’être pertinents, quelle que soit la méthode utilisée pour estimer cette probabilité.

III. Probability Ranking Principle PRP

Selon l'idée principale du PRP

- Pour obtenir le classement le plus efficace, il faut trier les documents selon leur **probabilité de pertinence** $P(R \mid \mathbf{d}, \mathbf{q})$.

- Le document d est jugé **pertinent** si sa probabilité d'appartenir à la classe des **pertinents est supérieure** à celle de **la classe non pertinente** :

$$\boldsymbol{P(R \mid d, q) > P(NR \mid d, q)}$$

Les documents sont triés selon le rapport de Odds : **RSV(q,d) = O(d)** $= \frac{\boldsymbol{P(R|d,q)}}{\boldsymbol{P(NR|d,q)}}$

Mais ce $P(R \mid \mathbf{d}, \mathbf{q})$ n'est pas directement observable :

- On ne connaît pas la distribution des documents pertinents à l'avance.
- On ne sait pas combien de documents pertinents existent dans la collection.
- On ne peut donc pas calculer $P(R \mid \mathbf{d}, \mathbf{q})$ directement.

IV. Du PRP vers un modèle probabiliste

Il faut donc **réécrire** $\boldsymbol{P(R \mid \mathbf{d}, \mathbf{q})}$ d'une manière qui relie la pertinence à des quantités **qu'on peut estimer** (fréquences de termes, présence/absence, etc.).

Appel a la loi de Bayes : sert a inverser une probabilité conditionnelle.

$$\boldsymbol{P(R \mid \mathbf{d}, \mathbf{q})} = \frac{P(d \mid \mathrm{R}, \mathrm{q})P(R \mid q)}{P(d \mid q)}. \text{ et } \boldsymbol{P(NR \mid \mathbf{d}, \mathbf{q})} = \frac{P(d \mid \mathrm{NR}, \mathrm{q})P(NR \mid q)}{P(d \mid q)}.$$

Substitution dans le rapport des deux probabilités

Remplaçons dans la définition du Odds :

$$O(d) = \frac{P(d \mid \mathrm{R}, \mathrm{q})\, P(R \mid q)/P(d \mid q)}{P(d \mid \mathrm{NR}, \mathrm{q}).\, P(NR \mid q)/P(d \mid q)}$$

Le dénominateur $P(d \mid q)$ est commun → il se simplifie :

$$O(d) = \frac{P(d \mid \mathrm{R}, \mathrm{q}) \cdot P(R \mid q)}{P(d \mid \mathrm{NR}, \mathrm{q}) \cdot P(NR \mid q)}$$

IV. Du PRP vers un modèle probabiliste

Simplification

Pour une même requête q, les probabilités a priori $P(R \mid q)$ et $P(NR \mid q)$
sont les des constantes et sont mêmes (elles ne dépendent pas du document d). Elles n'affectent pas l'ordre du classement, donc elles sont ignorées dans le calcul relatif :

$$O(d) = \frac{P(d \mid \mathrm{R}, \mathrm{q})}{P(d \mid \mathrm{NR}, \mathrm{q})}$$

Grâce à la loi de Bayes, on transforme un problème non observable (la pertinence réelle d'un document) en un problème estimable (probabilité d'observer un document selon qu'il soit pertinent ou non).

IV. Du PRP vers un modèle probabiliste

Question clé pour passer à un modèle probabiliste concret

Pour calculer $P(d \mid \mathrm{R}, \mathrm{q})$ et $P(d \mid \mathrm{NR}, \mathrm{q})$, il faut se poser 3 questions :

1. **Comment représenter le document d ?**
   Par exemple, comme un vecteur de termes binaires (présence/absence).

2. **Quelle distribution choisir pour $P(d \mid R)$ et $P(d \mid NR)$?**
   Probabilité d'observer le document sachant qu'il est pertinent ou non.
   Exemples : distribution binaire indépendante (BIM), Poisson, etc.

3. **Comment estimer les paramètres du modèle à partir de la collection et d'une requête q ?**
   Fréquence des termes, présence/absence, proportion de documents pertinents, etc.

V. BINARY INDEPENDENCE RETRIEVAL MODEL - BIR MODEL

Principes et Hypothèse

- Un **document** est vu comme un **ensemble d'événements**, chacun représentant la **présence ou l'absence d'un terme**.
- On représente un document d sous la forme d'un **vecteur binaire** :

$$d = (t_1, t_2, . , .. t_n)$$

où chaque t_i correspond à un terme du vocabulaire.

$$t_i = \begin{cases} 1, & \textit{si le terme } i \textit{ est présent dans le document} \\ 0, & \textit{sinon} \end{cases}$$

- L'événement "le mot $_{t_i}$est présent" est donc une **variable aléatoire de Bernoulli** (0 ou 1).

V. BINARY INDEPENDENCE RETRIEVAL MODEL - BIR MODEL

Principes et Hypothèse

- Hypothèse **d'indépendance** :
  la présence d'un terme dans un document est supposée **indépendante** de la présence des autres termes.

$$P(t_1, t_2, ..., t_n|R) = \prod_{i=1}^{n} P(t_i|R)$$

- Objectif : estimer la probabilité qu'un document soit **pertinent (R)** ou **non pertinent (NR)** pour une requête q.
- On classe les documents selon le rapport des probabilités (odds).

V. BINARY INDEPENDENCE RETRIEVAL MODEL - BIR MODEL

- Considérons un document comme une liste de termes
- Chaque terme de la requête est ensuite comparé à ceux du document pour estimer sa contribution à la **probabilité de pertinence**.
- Le modèle BIR cherche à estimer : $P(d \mid R)$ *et* $P(d \mid NR)$

  $P(d \mid R)$= probabilité d'observer le document s'il est **pertinent**

  $P(d \mid NR)$= probabilité d'observer le document s'il est **non pertinent**
- Sous l'**hypothèse d'indépendance des termes**, ces probabilités se factorisent :

$$P(d \mid R) = \prod_{i=1}^{n} P(t_i \mid R)$$

$$P(d \mid NR) = \prod_{i=1}^{n} P(t_i \mid NR)$$

$$\frac{P(d \mid R)}{P(d \mid NR)} = \prod_{i=1}^{n} \frac{P(t_i \mid R)}{P(t_i \mid NR)}$$

Donc: chaque mot agit indépendamment des autres dans le calcul de la probabilité globale du document.

V. BINARY INDEPENDENCE RETRIEVAL MODEL - BIR MODEL

- Chaque mot t_i est modélisé par une **loi de Bernoulli**, car il ne peut avoir que deux états : présent (1) ou absent (0) dans un document.

On définit alors :

- $p_i = P(t_i = 1 \mid R)$: probabilité qu'un terme apparaisse dans un document pertinent
- $q_i = P(t_i = 1 \mid NR)$: probabilité qu'un terme apparaisse dans un document non pertinent
- $1 - p_i = P(t_i = 0 \mid R)$: probabilité qu'il n'apparaisse pas dans un document pertinent.
- $1 - q_i = P(t_i = 0 \mid NR)$: probabilité qu'il n'apparaisse pas dans un document non pertinent.

Chaque terme t_i est représenté comme variable aléatoire x_i

$$P(d \mid R) = \prod_{i=1}^{n} P(t_i = x_i \mid R) = \prod_{i=1}^{n} p_i^{x_i} (1 - p_i)^{(1-x_i)}$$

$$P(d \mid NR) = \prod_{i=1}^{n} P(t_i = x_i \mid NR) = \prod_{i=1}^{n} q_i^{x_i} (1 - q_i)^{(1-x_i)}$$

V. BINARY INDEPENDENCE RETRIEVAL MODEL - BIR MODEL

$$O(d) = \frac{P(d|R)}{P(d|NR)} = \prod_{i=1}^{n} \frac{p_i^{x_i}(1-p_i)^{1-x_i}}{q_i^{x_i}(1-q_i)^{1-x_i}}$$

Problème :

- Ce rapport est un **produit de n termes**, chacun entre 0 et 1.
- Les produits de petits nombres → deviennent **très petits** → difficile à manipuler.
- De plus, les produits rendent l'interprétation et le tri compliqués.

Solution : Introduire le logarithme pour simplifier les produits

$$\log(ab) = \log a + \log b$$

Appliquer le logarithme , transformation du produit en somme :

$$\log O(d) = \sum_{i=1}^{n} \log \frac{p_i^{x_i}(1-p_i)^{1-x_i}}{q_i^{x_i}(1-q_i)^{1-x_i}}$$

V. BINARY INDEPENDENCE RETRIEVAL MODEL - BIR MODEL

$$\log O(d) = \sum_{i=1}^{n} \left[x_i \log \frac{p_i(1-q_i)}{q_i(1-p_i)} + \log \frac{1-p_i}{1-q_i} \right]$$

A ignorer parce que :

- Est une constante, Il ne contient pas x_i, donc il ne dépend pas du contenu du document.
- Il dépend uniquement des paramètres du modèle (p_i, q_i) qui sont fixés pour une requête donnée.

D'où la forme simplifiée du RSV :

$$RSV(q,d) = \sum_{i=1}^{n} x_i \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

Comment estimer p_i et q_i?

- **Avec données d'apprentissage.**
- **Sans données d'apprentissage.**

V. BINARY INDEPENDENCE RETRIEVAL MODEL - BIR MODEL

Estimation avec des données d'apprentissage

- On dispose d'un ensemble de documents jugés pertinents pour une requête (feedback utilisateur ou base étiquetée).
- On peut donc calculer empiriquement les probabilités.

Docs	pertinent (R)	non pertinent (NR)	**Total**
(ti = 1)	(r)	(n - r)	**(n)**
(ti = 0)	(R - r)	(N - n - R + r)	**(N - n)**
Total	**R**	**N - R**	**N**

$$p_i = P(t_i = 1 \mid R) = \frac{r}{R}$$

$$q_i = P(t_i = 1 \mid NR) = \frac{n - r}{N - R}$$

(r) nombre de documents pertinents contenant le terme (ti)

(n) nombre total de documents (pertinents ou non) contenant (ti)

(R) nombre total de documents pertinents (tous termes confondus)

(N) nombre total de documents dans la collection

V. BINARY INDEPENDENCE RETRIEVAL MODEL - BIR MODEL

Si je connais déjà quels documents sont pertinents… à quoi bon les estimer encore !!

- ✓ Quand on a un petit ensemble de documents **déjà jugés pertinents/non pertinents** (appelé *échantillon d'apprentissage* ou *feedback utilisateur*), on **apprend** les probabilités p_i et q_i pour chaque mot.
- ✓ Puis, on **applique ces probabilités** à **toute la collection**, pour classer les autres documents.
- ✓ Même si on connaît les documents pertinents dans un petit ensemble, on veut **prédire la probabilité de pertinence** des autres en fonction de leurs mots.

V. BINARY INDEPENDENCE RETRIEVAL MODEL - BIR MODEL

Estimation avec des données d'apprentissage

$$p_i = \frac{r}{R} \qquad \text{et} \qquad q_i = \frac{n-r}{N-R}$$

Substitution avec les estimateurs dans RSV

$$RSV(q,d) = \sum_{i=1}^{n} x_i \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

$$RSV(q,d) = \sum \log \frac{p(1-q)}{q(1-p)} =$$

$$= \sum \log \frac{\frac{r}{R} * \frac{N-n-R+r}{N-R}}{\frac{n-r}{N-R} * \frac{R-r}{R}} =$$

$$= \sum \log \frac{r/(R-r)}{(n-r)/(N-n-R+r)}$$

V. BINARY INDEPENDENCE RETRIEVAL MODEL - BIR MODEL

Estimation avec des données d'apprentissage

Formule avec lissage - Pour éviter les divisions par zéro, on ajoute 0.5 à chaque compteur :

$$RSV(q,d) = \sum_{t_i \in (d \cap q)} \log \frac{\frac{r_i + 0.5}{R - r_i + 0.5}}{\frac{(n_i - r_i + 0.5)}{(N - n_i - R + r_i + 0.5)}}$$

Extension BIR pondéré : Requêtes et documents sont pondérés

$$RSV(q,d_j) = \sum_{t_i \in (d \cap q)} w_{ij} * qtf_i * \log \frac{\frac{r_i + 0.5}{R - r_i + 0.5}}{\frac{(n_i - r_i + 0.5)}{(N - n_i - R + r_i + 0.5)}}$$

- Wij poids du terme i dans le document j
- qtfi poids du terme i dans la requête q

V. BINARY INDEPENDENCE RETRIEVAL MODEL - BIR MODEL

Estimation sans données d'apprentissage

On ne dispose d'aucune donnée d'apprentissage, c'est-à-dire :

- on ne sait pas quels documents sont pertinents pour la requête (pas de jugements humains),
- Donc on ne connaît pas ri (nb de docs pertinents contenant ti) ni R (nb total des docs pertinents)

Hypothèse de base (Croft & Harper, 1979

On applique une estimation a priori ou on suppose :

$$p_i = 0.5 \;\; et \;\; q_i = \frac{n_i}{N}$$

où : qi est estimé par la **fréquence du terme dans toute la collection**

Cette hypothèse revient à considérer qu'on n'a pas d'information de pertinence, donc $r_i = 0$.

V. BINARY INDEPENDENCE RETRIEVAL MODEL - BIR MODEL

Estimation sans données d'apprentissage

Formule du BIR :

$$RSV(q, d) = \sum_{i \in q \cap d} \log \frac{p_i(1 - q_i)}{q_i(1 - p_i)}$$

Remplaçons :

$$RSV(q, d) = \sum_{i \in q \cap d} \log \frac{0.5(1 - n_i/N)}{(n_i/N)(1 - 0.5)}$$

Simplifions :

$$RSV(q, d) = \sum_{i \in q \cap d} \log \frac{1 - n_i/N}{n_i/N} = \sum_{i \in q \cap d} \log \frac{N - n_i}{n_i}$$

V. BINARY INDEPENDENCE RETRIEVAL MODEL - BIR MODEL

Estimation sans données d'apprentissage

Formule avec lissage : Pour éviter les divisions par zéro (quand un mot est très rare ou absent), on ajoute 0.5 à chaque partie :

$$RSV(d,q) = \sum_{i \in q \cap d} \log \frac{N - n_i + 0.5}{n_i + 0.5}$$

Et c'est la version la plus utilisée en pratique.

V. BINARY INDEPENDENCE RETRIEVAL MODEL - BIR MODEL

Estimation sans données d'apprentissage

Extension BIR pondéré

$$RSV(q, d_j) = \sum_{i \in q \cap d_j} w_{ij} \cdot qtf_i \cdot \log \frac{N - n_i + 0.5}{n_i + 0.5}$$

Cela revient à une **pondération tf–idf probabiliste** :

- w_{ij} =importance du mot $t_i$dans le document d_j,
- qtf_i = poids du mot dans la requête,

V. BINARY INDEPENDENCE RETRIEVAL MODEL - BIR MODEL

Avantages

- Formalisation puissante et théoriquement fondée sur le principe du classement probabiliste (PRP).
- Permet une modélisation explicite de la notion de pertinence.
- Possibilité d'introduire une pondération des termes (en tenant compte de leur importance via les probabilités estimées).

Inconvénients

- Le modèle binaire de base ne tient pas compte directement de la fréquence des termes dans les documents (avant introduction de la pondération).
- Estimation des probabilités de pertinence difficile sans données d'apprentissage.
- Hypothèse d'indépendance entre les termes souvent critiquée,
  → les modèles prenant en compte cette dépendance n'apportent cependant pas d'amélioration significative en pratique.

VI. MODÈLE 2-POISSON [HARTER]

- Le modèle BIR basique ne tient pas compte de la fréquence des termes dans les documents.
- Il faut donc un modèle qui relie la fréquence d'un mot à sa probabilité de pertinence.
- L'idée : utiliser une loi statistique pour modéliser le nombre d'occurrences des mots dans les documents.

Les occurrences d'un mot dans un document sont distribuées de façon aléatoire: la probabilité qu'un mot apparaisse k fois dans un document suit une loi de Poisson :

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- k = nombre d'occurrences du terme dans un document
- λ = fréquence moyenne du terme dans les documents

VI. MODÈLE 2-POISSON [HARTER]

✓ Harter (1975) a remarqué que :

Les termes (mots) ne sont pas distribués selon une loi de Poisson dans tous les documents – Les mots qui traitent le sujet du document ont une distribution différente de ceux apparaissent de manière marginale dans le document.

✓ Il propose une combinaison de deux lois de Poisson :

- Une pour les documents élites (ceux qui traitent du sujet du terme)
- Une pour les documents non élites (où le mot apparaît par hasard)

D'où le nom "2-Poisson".

VI. MODÈLE 2-POISSON [HARTER]

Formulation mathématique du modèle 2-Poisson

La fréquence du terme t dans un document suit une distribution mixte :

$$P(t=k) = P(E)\frac{(\lambda_1)^k e^{-\lambda_1}}{k!} + P(\neg E)\frac{(\lambda_0)^k e^{-\lambda_0}}{k!}$$

où :

- $P(E)$:probabilité qu'un document soit "élite"
- λ_1 :moyenne des fréquences du mot dans les documents élites
- λ_0 :moyenne des fréquences dans les documents non élites

Le modèle **2-POISSON** est **mathématiquement complexe,** il faut estimer trois paramètres pour chaque mot. En pratique, on n'a pas assez de données pour estimer correctement ces valeurs.

Mais il a inspiré la forme du modèle BM25 : qui a introduit la pondération par fréquence de terme (tf), et la normalisation par la longueur du document, sans passer explicitement par une loi de Poisson.

VII. MODÈLE BM25

De la loi de Poisson à BM25

Étendre le modèle probabiliste pour intégrer la fréquence des termes et la notion d'élite

Pour relier cette idée au modèle probabiliste de la RI, on introduit :

- $p_i = P(E \mid R)$:probabilité qu'un document pertinent soit élite pour le terme t_i
- $q_i = P(E \mid NR)$: probabilité qu'un document non pertinent soit élite pour le terme t_i

Ainsi :

$$P(t_i = k \mid R) = P(E \mid R)\frac{e^{-\lambda_1}\lambda_1^k}{k!} + P(\neg E \mid R)\frac{e^{-\lambda_0}\lambda_0^k}{k!}$$

$$P(t_i = k \mid NR) = P(E \mid NR)\frac{e^{-\lambda_1}\lambda_1^k}{k!} + P(\neg E \mid NR)\frac{e^{-\lambda_0}\lambda_0^k}{k!}$$

Cela permet de **lier la fréquence observée** du mot t_i à la **probabilité de pertinence**.

VII. MODÈLE BM25

Le modèle **BM25**, proposé par **Robertson et Walker (1994)**, est une version simplifiée et plus pratique du **modèle à deux Poisson**. Il vise à estimer plus efficacement la pertinence d'un document pour une requête.

Évolutions principales :

- Le modèle remplace la loi de Poisson par **une fonction plus simple**, qui relie la fréquence d'un mot à son influence sur la pertinence.
- Il introduit une **correction selon la longueur du document**, pour éviter d'avantager les textes longs.
- Il conserve le **poids IDF probabiliste**, qui donne plus d'importance aux mots rares dans la collection.

BM25 combine la fréquence d'un mot, la longueur du document et la rareté du terme pour estimer la pertinence de façon équilibrée et efficace.

VII. MODÈLE BM25

Forme finale du modèle BM25

$$RSV_{BM25}(q,d) = \sum_{i \in q} \log \frac{N - n_i + 0.5}{n_i + 0.5} \times \frac{(k_1 + 1)\, tf_i}{k_1 \left((1-b) + b\frac{dl}{avdl}\right) + tf_i}$$

N	Nombre total de documents
ni	Nombre de documents contenant le terme (ti)
tfi	Fréquence du terme (ti) dans le document
dl	Longueur du document
avdl	Longueur moyenne des documents
k1	Contrôle la saturation de la fréquence (1.2–2.0 typiquement)
b	Contrôle la normalisation par longueur (souvent ≈ 0.75)

EXERCICE

Fait en cours