4. Modèles probabilistes

Attention document en version préliminaire - VERSION non définitive

Introduction

Dans ce chapitre nous présentons la famille de modèles de recherche dits *probabilistes*. L'objectif de ces modèles, introduits dans les années 60, est d'apporter une explication probabiliste au problème de la recherche d'information. La recherche d'information suit un procédé imprécis et incertain. En effet, l'expression des besoins d'un utilisateur par une requête reste une représentation imprécise de ses besoins. Ainsi la théorie des probabilités semble être un cadre adapté à la gestion de ces problèmes d'imprécision et d'incertitude. Le principe général des modèles probabilistes consiste à inférer, étant données d'une part la représentation d'un document et d'autre part la représentation d'une requête, la probabilité de pertinence du document sachant la requête. L'idée principale est que plus la probabilité de pertinence d'un document de la collection est élevée plus le document répond au besoin de l'utilisateur tel que formulé par sa requête. Ainsi les probabilités sont un outils pour l'ordonnancement des documents par ordre de pertinence. La majorité des méthodes associées aux modèles probabilistes reposent sur ce principe d'ordonnancement. Nous commencerons par la présentation de ce dernier en section 4.1. Nous introduirons ensuite le *modèle d'indépendance binaire* en section 4.2 suivie du *modèle 2-Poisson* en section 4.3. Pour finir, nous présenterons le modèle *OKAPI BM25* en section 4.4.

Mots-clés

- Probabilité de pertinence, principe d'ordonnancement probabiliste
- Modèle d'Indépendance Binaire (MIB), modèle 2-Poisson, modèle OKAPI BM25, paramètre d'un modèle de recherche, hypothèses du modèle MIB
- Fonction de décision d'ordonnancement ou critère d'ordonnancement probabiliste (RSV), score RSV, mesure RSV
- maximum de vraisemblance

4.1 Le principe d'ordonnancement probabiliste

Le point de départ des modèles probabilistes est un principe d'ordonnancement, connu comme le **principe d'ordonnancement probabiliste** proposé par Roberston et Spärck Jones en 1977 sous la formulation suivante :

*"If a reference retrieval system's response to each request is a **ranking** of the **documents** of the collection **in order of decreasing probability of relevance** to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data." (Probability Ranking Principle)*

Pour simplifier, en reprenant la formulation proposée par Spärck Jones, le modèle de recherche probabiliste peut être résumé par la mise en place d'un système pour répondre à la question basique suivante :

- **Quelle est la probabilité qu'un utilisateur juge *ce* document pertinent pour *cette* requête ?**

Cette question implique deux hypothèses sur la pertinence :

1. La pertinence est un attribut binaire : le document est soit pertinent, soit non pertinent.
2. La pertinence de chaque document ne dépend pas de la pertinence des autres documents.

4.1.1 Une expérience aléatoire support aux modèles probabilistes

En termes statistiques, cette question consiste donc à estimer la probabilité d'un jugement de pertinence positif associé à un couple document/requête (d,q) donné. Plus formellement, afin de construire les modèles probabilistes nous permettant de calculer cette probabilité, nous introduisons comme support **une expérience aléatoire** où une paire (d,q) est une réalisation de cette expérience. L'expérience aléatoire est la suivante :
"Tirer simultanément (avec remise)[1] d'une urne contenant à la fois un ensemble prédéfini de documents d_i et un ensemble prédéfini de requêtes q_l, un couple (d,q)."
Cette réalisation de l'expérience est alors liée soit à l'événement "le document d est pertinent pour la requête q" soit à l'événement contraire "le document d n'est pas pertinent pour la requête q". Nous introduisons donc la variable aléatoire binaire R telle que l'**événement** $\{R=1\}$ **(resp.** $\{R=0\}$**) correspond à "le document** d **est pertinent pour la requête** q**" (resp. "le document** d **n'est pas pertinent pour la requête** q**")**[2]. Plus précisément, nous introduisons aussi la variable D qui a pour réalisation $\{D=d\}$ **(le document** d **a été tiré simultanément à requête** q**)** et la variable Q qui a pour réalisation $\{Q=q\}$ **(la requête tirée simultanément au document** d **est** q**)**. Le cadre de cette expérience aléatoire nous permet de formaliser mathématiquement la question précédente par le calcul de la probabilité suivante :

Definition 4.1.1 — Probabilité de pertinence. Soient D et Q les variables aléatoires ayant pour réalisation le fait de tirer simultanément un document d et une requête q. Soit R une variable aléatoire binaire de pertinence ou non pertinence du document d pour la requête q. Nous formulons, la probabilité de pertinence par

$$P(R=1|D=d,Q=q) \tag{4.1}$$

où l'événement $\{R=1\}$ signifie que lors de la réalisation de l'expérience aléatoire, le document d observé est pertinent pour la requête q observée.

[1] Nous retrouvons ici l'hypothèse 2
[2] Nous retrouvons ici l'hypothèse 1

à savoir quelle est la probabilité que la pertinence soit positive sachant que j'ai observé l'événement $\{D = d\}$ et l'événement $\{Q = q\}$ (le document d et la requête q sont observés lors de la réalisation de l'expérience) ?

Les modèles probabilistes que nous allons présenter sont chacun une modélisation de l'expérience aléatoire introduite plus haut. Les paramètres de ces modèles seront alors à estimer. Enfin l'objectif principal de ces modélisations sera de permettre, pour toute nouvelle requête q', l'ordonnancement ou le tri des documents d en fonction de la probabilité (a posteriori) $P(R = 1|D = d, Q = q')$.

4.1.2 Simplification de notations

Par simplification on note souvent d pour l'événement $\{D = d\}$ et q pour $\{Q = q\}$ et on doit donc estimer la probabilité :

$$P(R = 1|d, q)$$

De même, on note souvent R pour $\{R = 1\}$ (pertinent) et NR pour $\{R = 0\}$ (non pertinent). Ces événements étant disjoints et les seuls événements possible de l'expérience, on a donc d'après la formule des probabilités totales :

$$P(R|d, q) = 1 - P(NR|d, q)$$

Enfin, dans toute la suite de chapitre, nous considérons que la requête q est donnée donc connue. Pour simplifier les notations, nous n'écrirons plus q dans le conditionnement mais il faut garder à l'esprit sa présence. Ainsi, nous écrirons $P(R|d)$ au lieu de $P(R|d, q)$, etc.

4.1.3 Astuce de calcul et critère d'ordonnancement RSV

Considérons que q est donné et il faut estimer $P(R|d)$ ce qui au premier abord peut ne pas sembler évident. Par contre, à partir d'informations sur l'ensemble de documents pertinents, il peut être plus simple de calculer $P(d|R)$. Alors une astuce est d'appliquer la règle de Bayes afin de calculer la probabilité de pertinence.

C'est l'idée énoncée par Van Rijsbergen de la manière suivante :
For a given query, if we know some documents that are relevant, terms that occur in those documents should be given greater weighting in searching for other relevant documents.
By making assumptions about the distribution of terms and applying Bayes Theorem, it is possible to derive weights theoretically.
En appliquant le théorème de Bayes, on obtient une nouvelle formulation de la probabilité de pertinence.

Theorem 4.1.1 — Theorem de Bayes et Probabilité de pertinence. Selon la règle de Bayès, nous pouvons formuler la probabilité de pertinence par

$$P(R|d) = \frac{P(d|R)P(R)}{P(d)} \tag{4.2}$$

où $P(R)$ est la probabilité de pertinence a priori (probabilité de trouver un document pertinent ou probabilité de pertinence d'un document quelconque), $P(d|R)$ est la probabilité que si un document pertinent est trouvé alors il s'agit de d (ou probabilité que d fasse partie des documents pertinents), enfin $P(d)$ est la probabilité que d soit choisi. $P(d)$ est aussi nommée "constante de normalisation".

et de manière similaire, pour la probabilité de non pertinence, on a :

$$P(NR|d) = \frac{P(d|NR)P(NR)}{P(d)}.$$

Aussi, la formule des probabilités totales nous permet d'avoir :

$$P(R|d) + P(NR|d) = 1$$

Alors, l'idée sur laquelle repose le principe d'ordonnancement probabiliste est la règle de décision suivante :

Proposition 4.1.2 — Règle de décision. Un document d est sélectionné comme pertinent vis-à-vis d'une requête q si

$$P(R|d) > P(NR|d) \tag{4.3}$$

Par conséquent, une façon de trier les documents est de les ordonner selon les valeurs prises par la statistique suivante :

$$RSV(d,q) = \frac{P(R|d)}{P(NR|d)}$$

Cette statistique correspond au score du niveau de pertinence du document d pour la requête q. Plus le score sera élevé, plus le document sera bien classé en terme de pertinence pour la requête q.

Cependant, nous avion relevé précédemment le fait qu'il est plus évident de calculer $P(d|R)$ plutôt que $P(R|d)$. Alors, en réutilisant la règle de Bayès, nous obtenons l'expression équivalente suivante :

$$RSV(d,q) = \frac{P(d|R)}{P(d|NR)} \frac{P(R)}{P(NR)}.$$

Or, $\frac{P(R)}{P(NR)}$ est une constante positive indépendante du document. Nous pouvons donc nous passer de son calcul puisque notre tâche d'intérêt est le classement des document par niveau de pertinence vis-à-vis d'une requête q. Tous ces éléments, nous permette enfin, d'énoncer la fonction score du principe d'ordonnancement probabiliste.

Definition 4.1.2 — Score RSV (Retrieval Status Value) : critère d'ordonnancement probabiliste. Soient une requête q et un document d, RSV est le score de pertinence associé au document d vis-à-vis de la requête q. Ce score est définie par

$$RSV(d,q) = \frac{P(d|R)}{P(d|NR)} \tag{4.4}$$

où $P(d|R)$ est la probabilité que d fasse partie des documents pertinents et $P(d|NR)$ est la probabilité que d fasse partie des documents non pertinents.

Ainsi, pour une requête q fixée, le calcul de ce score RSV pour différents documents d va permettre de les ordonner par niveau de pertinence selon le principe d'ordonnancement probabiliste. Se pose alors la question de l'estimation des probabilités permettant le calcul de ce score. Plusieurs types de solutions sont envisageables. Nous allons en présenter deux dans les sections suivantes : celle proposée par le modèle d'indépendance binaire et celle des modèles 2-poisson et OKAPI BM25.

4.2 Modèle d'indépendance binaire (MIB)

Le modèle d'indépendance binaire est associé au principe d'ordonnancement probabiliste et repose sur 4 grandes hypothèses :

1. **H1 (hypothèse de représentation vectorielle binaire de d et q)** : **Un document et respectivement une requête sont représentés comme un ensemble d'évènements t_j dénotant la présence ou l'absence d'un terme dans le document, respectivement la requête**. **Les termes t_j représentent le vocabulaire de la collection**. Un document d et respectivement une requête q seront donc représentés comme un vecteur de caractéristiques binaires $\mathbf{d} = (t_1, \cdots, t_j, \cdots, t_V)$ avec $\{t_j = 0\}$ ou $\{t_j = 1\}$ qui indiquent respectivement la présence ou l'absence du $i$ème terme d'index t_j.
2. **H2 (hypothèse d'indépendance des termes)** : Les termes apparaissent dans les documents et dans les requêtes de manière **indépendante** : les $\{t_j = 0 \text{ ou } 1\}$ sont mutuellement indépendants. Cette hypothèse est aussi connue sous le nom d'hypothèse de *Naïve Bayes*.
3. **H3 (hypothèse de décision et d'ordonnancement probabiliste)** : Pour une requête q fixée, la règle de décision sur la pertinence d'un document et la statistique d'ordonnancement des documents sont respectivement celles définies en (4.1.2) et en (4.1.2) dans la section précédente. Plus le score RSV est élevé pour un document d, plus le document est considéré comme pertinent par rapport à la requête q. Il joue le même rôle que la fonction de score des modèles vectoriels.
4. **H4 (hypothèse de répartition uniforme des termes absents)** : Les termes non présents dans la requête sont **uniformément répartis** dans les documents pertinents et non pertinents.

R L'hypothèse 4 signifie qu'on suppose que les termes non présents dans la requête ont autant de chance de se retrouver dans des documents pertinents comme dans des documents non pertinent. L'idée sous-jacente est que l'on va considérer les termes de la requêtes et ignorer les termes non présents dans requête. On suppose que ces derniers n'ont aucune influence.

Ces différentes hypothèses vont être utilisées pour estimer le score de pertinence RSV (4.1.2). On va notamment s'intéresser à la probabilité $P(d|R)$ (pour rappel, $P(D = d|R = 1, Q = q)$).

4.2.1 Le critère d'ordonnancement RSV sous les hypothèse de MIB

- En prenant en compte **H1**, les probabilités conditionnelles de d sont assimilées au probabilités conditionnelles de son vecteur binaire représentatif $\mathbf{d} = (t_1, ...t_V)$. Ainsi,

  $$P(d|R) = P(\mathbf{d}|R) = P(\mathbf{d} = (t_1, \cdots, t_V)|R)$$

  avec $t_1, \cdots, t_V$ les termes du vocabulaire.
- Puis, d'après l'hypothèse **H2** d'indépendance mutuelle entre les termes d'un document et d'une requête, on a :

  $$P(\mathbf{d}|R) = \prod_{t_j \in \mathbf{d}} P(t_j|R) \prod_{t_j \notin \mathbf{d}} (1 - P(t_j|R))$$

  où $P(t_j|R)$ est la probabilité que le terme t_j apparaisse dans un document pertinent ou probabilité de pertinence du terme t_j et $(1 - P(t_j|R))$ est la probabilité que le terme t_j n'apparaisse pas dans un document pertinent. $P(\mathbf{d}|R)$ peut être estimée comme étant le produit des probabilités de pertinence associées à chaque terme dans le document, multiplié par le produit des probabilités que les termes absents n'apparaissent pas dans un document pertinent.

  R Remarquons que de manière similaire, nous avons :

  $$P(\mathbf{d}|NR) = \prod_{t_j \in \mathbf{d}} P(t_j|NR) \prod_{t_j \notin \mathbf{d}} (1 - P(t_j|NR)).$$

t_j une variable de Bernouilli

Notons que t_j peut être vu comme une variable aléatoire suivant la loi de Bernouilli. En effet, nous avons :

$$\mathbf{d} = (t_1 = x_1, \cdots, t_V = x_V)$$

tel que,

$$x_j = \begin{cases} 1 \text{ si le terme est présent} \\ 0 \text{ sinon.} \end{cases}$$

Par soucis de simplification de notations, si nous notons :
- $p_j = P(t_j = 1|R)$ la probabilité de pertinence de t_j
- $s_j = P(t_j = 1|NR)$ la probabilité de non-pertinence de t_j

nous obtenons une écriture explicite des distributions :

$$P(\mathbf{d}|R) = \prod_{i=j,t_j=1}^{V} p_j \prod_{j=1,t_j=0}^{V} (1-p_j) \tag{4.5}$$

$$= \prod_{j=1}^{V} p_j^{t_j} (1-p_j)^{(1-t_j)} \tag{4.6}$$

et de manière similaire,

$$P(\mathbf{d}|NR) = \prod_{j=1}^{V} s_j^{t_j} (1-s_j)^{(1-t_j)} \tag{4.7}$$

R On a aussi $1-p_j = P(t_j = 0|R)$ et $1-s_j = P(t_j = 0|NR)$.

Avec les écritures de probabilités de pertinence et non pertinence (4.2.1) et (4.2.1), on aboutit à la réécriture suivante du score de pertinence :

$$RSV(d,q) = \prod_{j=1}^{V} \left(\frac{p_j}{s_j}\right)^{t_j} \left(\frac{1-p_j}{1-s_j}\right)^{(1-t_j)} \tag{4.8}$$

- Selon **H4**, **les termes non présents dans la requête sont uniformément répartis dans les documents pertinents et non pertinents**. Cela signifie que si on considère une requête q comme un vecteur $\mathbf{q} = (r_1, \cdots, r_j, \cdots, r_V)$ avec $r_j = 0$ ou 1 qui indique la présence ou l'absence du terme d'index r_j dans la requête q alors nous avons pour $r_j = 0$ la propriété $p_j = s_j$. C'est à dire qu'un terme absent dans la requête a "autant de chance" d'apparaître dans un document pertinent que dans un document non pertinent pour la requête. On a donc

$$RSV(d,q) = \prod_{j:t_j=r_j=1} \left(\frac{p_j}{s_j}\right) \prod_{j:t_j=0,r_j=1} \left(\frac{1-p_j}{1-s_j}\right) \tag{4.9}$$

Par conséquent, **on ne considère dans les produits que les termes qui apparaissent dans la requête**[3]. Si on s'intéresse à l'équation précédente (4.2.1), le premier produit correspond aux termes de la requête trouvés dans le document (événement $\{r_j = 1\}$ et événement

[3] Ce qui allège les calculs en pratique.

$\{t_j = 1\}$) et le deuxième produit correspond aux termes de la requête non trouvée dans le document (événement $\{r_j = 1\}$ et événement $\{t_j = 0\}$).
On multiplie le terme de droite par $\prod_{i:t_i=r_i=1} \left(\frac{1-p_i}{1-s_i} \times \frac{1-s_i}{1-p_i}\right) = 1$ ce qui revient à ajouter les termes de la requête trouvés dans le document ($\{t_j = 1\}$ et $\{r_j = 1\}$), on a donc

$$RSV(d,q) = \prod_{j:t_j=r_j=1} \left(\frac{p_j(1-s_j)}{s_j(1-p_j)}\right) \prod_{j:r_j=1} \left(\frac{1-p_j}{1-s_j}\right)$$

Le second produit concerne tous les termes de la requête et est constant pour une requête donnée. De plus ce second produit est constant quelque soit le document. Il est donc inutile pour l'ordonnancement. Ainsi, seule la quantité à estimer pour ordonner les documents est :

$$RSV(d,q) = \prod_{j:t_j=r_j=1} \left(\frac{p_j(1-s_j)}{s_j(1-p_j)}\right)$$

En pratique on prend le logarithme de ce terme. Cette transformation ne change pas l'ordre des scores puisque la fonction logarithme est strictement croissante sur $\mathbb{R}_+^*$. Par contre, passer au logarithme présente l'avantage de fournir une somme et non un produit de termes ce qui permet de traiter plus facilement les cas où les quantités manipulées sont très petites.

Le critère d'ordonnancement du modèle d'indépendance binaire, nommé $RSV(d,q)$ se réécrit donc

$$RSV(d,q) = \log \prod_{j:t_{ij}=r_j=1} \frac{p_j(1-s_j)}{s_j(1-p_j)} = \sum_{j:t_j=r_j=1} \log \left(\frac{p_j(1-s_j)}{s_j(1-p_j)}\right)$$

Il s'agira donc d'ordonner les documents suivant la valeur décroissante de la mesure RSV.

Definition 4.2.1 La formulation finale du **critère RSV d'ordonnancement du Modèle d'Indépendance Binaire** est :

$$RSV^{MIB}(d,q) = \sum_{j:t_j=r_j=1} log \left(\frac{p_j(1-s_j)}{s_j(1-p_j)}\right) \tag{4.10}$$

où $r_j = 1$ (resp. $t_j = 1$) signifie la présence du $j$ème terme du vocabulaire dans la requête q (resp. dans le document d).

R Une autre écriture possible de $RSV^{MIB}(d,q)$ est la suivante :

$$RSV^{MIB}(d,q) = \sum_{j:t_j=r_j=1} \left(log \left(\frac{p_j}{1-p_j}\right) + log \left(\frac{1-s_j}{s_j}\right)\right).$$

Pour pouvoir calculer ce score, il nous reste à déterminer une façon de calculer ses paramètres : ici les probabilités p_j et s_j.

4.2.2 Estimation des probabilités et paramètres du score

Pour tous les termes t_j de la requête, les probabilités p_j et s_j sont estimées d'après le principe du **maximum de vraisemblance** en utilisant une collection de documents $\mathscr{C}$ et une base de requêtes pour lesquelles on connaît les documents pertinents et non pertinents dans $\mathscr{C}$.

Considérons le tableau (4.1) qui représente la table de contingence des occurrences des documents dans la collection pour une requête donnée. On note,

- N le nombre total de documents dans la collection
- R_p le nombre total de documents pertinents (pour la requête)
- df_{t_j} le nombre de documents contenant t_j
- r le nombre de documents pertinents contenant t_j

	Documents	Pertinent (R)	Non-Pertinent (NR)	Total
Terme présent	$\{t_j = 1\}$	r	$df_{t_j} - r$	df_{t_j}
Terme absent	$\{t_j = 0\}$	$R_p - r$	$N - df_{t_j} - R_p + r$	$N - df_{t_j}$
	Total	R_p	$N - R_p$	N

Table 4.1: Table de contingence des occurrences des documents dans la collection pour une requête donnée.

D'après le tableau de contingence (4.1), on obtient les **formules d'estimations théoriques** des probabilités p_j et s_j suivantes :

- $p_j = \frac{r}{R_p}$
- $s_j = \frac{df_{t_j} - r}{N - R_p}$

aussi,

- $1 - p_j = \frac{R_p - r}{R_p}$
- $1 - s_j = \frac{N - df_{t_j} - R_p + r}{N - R_p}$.

En injectant ces estimations des probabilités dans l'équation (4.2.1), on en déduit l'estimation du critère $RSV^{MIB}(d,q)$ comme suit :

$$\widehat{RSV}^{MIB}(d,q) = \sum_{j:t_j=r_j=1} log\left(\frac{r/(R_p - r)}{(df_{t_j} - r)/(N - df_{t_j} - R_p + r))}\right) \tag{4.11}$$

Enfin, puisque une probabilité prend ses valeur sur l'intervalle $[0,1]$, nous lissons les probabilités en ajoutant 0.5 à chaque quantité du tableau. Cette astuce permet d'éviter les zéros. On obtient alors l'estimation finale suivante du score du critère RSV :

$$\widehat{RSV}^{MIB}(d,q) = \sum_{j:t_j=r_j=1} log\left(\frac{(r+0.5)/(R_p - r + 0.5)}{(df_{t_j} - r + 0.5)/(N - df_{t_j} - R_p + r + 0.5))}\right) \tag{4.12}$$

Ainsi, pour calculer le score de pertinence RSV, il suffit de parcourir les termes de la requête et sommer uniquement sur les termes de la requête présents dans les documents de la collection.

R Attention : pour chaque $j$ème terme de la requête (indice de sommation) les valeurs de r, R_p et de df_{t_j} peuvent varier. Leur valeur dépend de l'indice $j \in \{1, \cdots, V\}$.

R En pratique, sans données de retour de l'utilisateur, nous n'avons pas accès à des données labalisées (document pertinent/non pertinent). Alors dans ce cas, il n'est pas possible d'utiliser l'approche théorique de calcul des probabilités fournie par la table de contingence. Sans ces informations, il est possible d'approximer s_j par IDF (c'est à dire $\frac{df_{t_j}}{N}$). Quant à p_j plusieurs possibilités d'approximation existent :

- considérer $p_j = P(t_j|R)$ constante, par exemple : 0.5 si aucune information n'est disponible [Croft and Harper]
- considérer p_j proportionnelle à la probabilité d'occurrence dans la collection
- considérer p_j proportionnelle au logarithme de la probabilité d'occurrence dans la collection [GREIFF, Sigir 98]

4.2.3 Avantages et inconvénients du modèle MIB

Le modèle MIB présente donc les avantages d'une formalisation puissante et d'une modélisation explicite de la notion de pertinence. Cependant des inconvénients viennent les contre balancer. Par exemple la fréquence des termes n'est pas prise en compte, sans donner d'apprentissage il est difficile d'estimer les probabilités p_j et s_j. De plus, l'hypothèse d'indépendance mutuelle des termes **(H2)** est souvent critiquée.

4.3 Modèle 2-Poisson

Parmi les modèles probabilistes, certains vont plus loin que le modèle MIB en considérant des probabilités d'occurrences de mot. Il semble par exemple naturel de penser qu'un terme peu fréquent dans un document a peu de chance de bien représenter le document et que par conséquent, il y a peu de chance qu'un utilisateur désirant ce document utilise ce terme dans sa requête. Ceci motive l'idée d'introduire des modèles fondés sur la distribution des mots dans un document. Par exemple, le modèle 2-Poisson considère que les occurrences d'un mot suivent une loi de Poisson. Nous allons voir qu'en distinguant les documents de la collection en 2 catégories alors il sera possible d'envisager que les occurrences de mot suivent un mélange de deux lois de Poisson différentes, d'où l'intitulé du *modèle 2-Poisson*.

4.3.1 Distribution aléatoire des occurrences d'un terme

L'idée de base est que les occurrence d'un mot dans un document sont distribuées de manière aléatoire. Plus précisément, le probabilité qu'un mot apparaisse $tf_{t_j,d}$ fois dans un document suit une loi de probabilité. Dans le cas du modèle 2-Poisson, cette loi est une loi de Poisson. Cette loi de probabilité décrit le comportement du nombre d'évènements se produisant dans un laps de temps fixé sous la condition que ces événements se produisent avec une fréquence moyenne connue et indépendante du temps écoulé depuis l'événement précédent. Dans notre contexte, l'événement est "le terme t_j apparaît $tf_{t_j,d}$ fois dans le document d" et est noté $\{TF_j = tf_{t_j,d}\}$. On dit que TF_j est une variable aléatoire de loi de Poisson de paramètre $\lambda > 0$. Elle est discrète et à valeur dans $\mathbb{N}$. On écrit $TF_j \sim \mathscr{P}(\lambda)$.

Definition 4.3.1 — Loi de Poisson. La loi de Poisson de la variable TF_j a pour distribution :

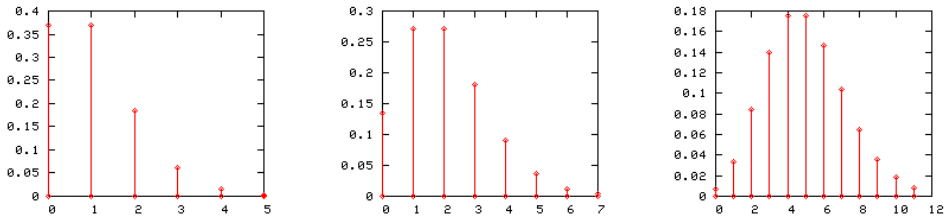
$$P(TF_j = tf_{t_j,d}) := \lambda^{tf_{t_j,d}} \frac{e^{-\lambda}}{tf_{t_j,d}!} \tag{4.13}$$

pour $tf_{t_j,d} \in \mathbb{N}$ et $\lambda > 0$. λ correspond à la moyenne de la variable TF_j.

C'est à dire que le paramètre λ de la loi de Poisson correspond à la moyenne des occurrences du terme t dans le document d.

R Notons que pour deux valeurs différentes du paramètre λ, nous sommes en présence de deux lois de Poisson différentes et donc de deux distributions différentes. Voici par exemple respectivement trois illustrations[4] graphiques de trois lois de Poisson de paramètres respectifs $\lambda = 1, 2, 5$.

[4] Empruntées à wikipédia : `https://fr.wikipedia.org/wiki/Loi_de_Poisson`.



4.3.2 Notion d'élitisme

Dans un objectif de distinguer les lois (ou distributions) des termes dans les documents où ce terme est représentatif de ceux où il ne l'est pas, nous définissons la notion d'**élitisme**.

Definition 4.3.2 — Notion d'élitisme. Nous considérons deux groupes de documents :

- E le groupe élite, où les documents traitent du thème représenté par le terme t. C'est à dire dans lesquels le terme t sera le plus fréquent.
- $\bar{E}$ le groupe complémentaire à E contenant les documents ne traitant pas du thème représenté par le terme t. C'est à dire dans lesquels l'occurrence de t est marginale.

Le principe des modèle probabilistes utilisant cette notion d'élitisme est fondé sur le fait de considérer que les lois du terme t dans les deux groupes E et $\bar{E}$ sont différentes. Ceci permet d'écrire la loi des termes dans les documents comme un mélange de deux lois de Poissons : on dit qu'un terme suit une distribution mixte 2-Poisson.

Definition 4.3.3 — Distribution mixte 2-Poisson. Soit $P(E)$ la probabilité a priori que le document soit élite, nous définissons la loi (ou distribution) mixte 2-Poisson comme suit :

$$P(TF_j = tf_{t_j,d}) := P(E)\lambda_1^{tf_{t_j,d}} \frac{e^{-\lambda_1}}{tf_{t_j,d}!} + P(\bar{E})\lambda_2^{tf_{t_j,d}} \frac{e^{-\lambda_2}}{tf_{t_j,d}!} \tag{4.14}$$

où

- λ_1 et λ_2 correspondent respectivement aux paramètres de moyenne des occurrences des termes dans les documents élites (qui appartiennent au groupe E) et non élites (qui appartiennent au groupe $\bar{E}$).
- $\lambda_1 \leq \lambda_2$

R Pour les mots vides, la notion d'élitisme n'a pas lieu d'être. La distribution des mots vides est supposée uniforme sur tous les documents.

4.3.3 Indexation probabiliste

Le principe de l'indexation probabiliste repose sur le fait qu'on souhaite sélectionner un terme t_j pour représenter un document si ce terme apparaît plus fréquemment dans ce document que dans un autre choisi au hasard. L'objectif est de distinguer les distributions des termes dans les documents où ces termes sont représentatifs de ceux dans lesquels ils ne le sont pas. Pour ce faire, on utilise l'équation (4.3.3) et on cherche à estimer les paramètres λ_1 et λ_2 :

- Si les estimations donnent des valeurs proches, on est en présence d'un terme peu spécifique
- Si les estimations donnent des valeurs différentes, on est en présence d'un terme spécifique qu'il faut utiliser pour décrire les documents élites de ce terme

Pour mesurer le degré de recouvrement des deux lois de poissons on calcule :

$$z = \frac{\lambda_1 - \lambda_2}{\sqrt{\lambda_1 + \lambda_2}}$$

Puis on décide qu'un terme t_j doit indexer un document d si $\eta > 0$ avec

$$\eta = P(d \in E | TF_j = tf_{t_j,d}) + z$$

où

- η est utilisée comme pondération associée au terme t pour le document d
- $P(d \in E | TF_j = tf_{t_j,d})$ est estimée en utilisant le théorème de Bayes :

$$\begin{aligned} P(d \in E | TF_j = tf_{t_j,d}) &= \frac{P(TF_j = tf_{t_j,d} | d \in E) P(d \in E)}{P(TF_j = tf_{t_j,d})} \\ &= \frac{\pi . e^{-\lambda_1} \lambda_1^{tf_{t_j,d}}}{\pi . e^{-\lambda_1} \lambda_1^{tf_{t_j,d}} + (1-\pi) . e^{-\lambda_2} \lambda_2^{tf_{t_j,d}}} \end{aligned}$$

où π est la probabilité a priori qu'un document d soit dans le groupe élite (i.e. $P(E)$).

4.4 Modèle OKAPI BM 25

Le modèle OKAPI BM 25, aussi nommé BM 25, est une référence dans le développement des systèmes de recherche. Il se fonde sur un "principe d'indexation probabiliste" dont l'idée sous-jacente est qu'un bon descripteur de document est un terme assez fréquent dans ce document mais qui est relativement rare dans la collection. Cette idée a pour notions de bases la division des documents en deux ensembles élites E et non élites $\bar{E}$ et aussi celles de probabilités de pertinence d'un terme p_j et s_j. En effet, sous le modèle BM 25 nous allons intégrer la notion d'élitisme dans le calcul de ces probabilités de pertinence ce qui va induire une réécriture du critère d'ordonnancement propre au BM 25.

4.4.1 Intégration de la notion d'élitisme dans le modèle probabiliste de base

Le modèle BM 25 est fondé sur le critère RSV^{MIB} décrit par l'équation (4.2.1). Il intègre dans cette équation

- la notion d'élitisme des termes, c'est à dire la loi mixte 2-Poisson des fréquences des termes dans les documents
- L'hypothèse que la fréquence d'un mot dans un document ne dépend que de l'appartenance du document à l'ensemble élite

En particulier, cela signifie qu'on va considérer que,

$$\begin{cases} p_j = P(TF_j = tf_{t_j,d} | R) \\ s_j = P(TF_j = tf_{t_j,d} | NR) \end{cases} \tag{4.15}$$

avec $tf_{t_j,d}$ le nombre d'occurrences du terme dans le document. Quand le terme est absent $tf_{t_j,d} = 0$. Ainsi, quand le terme est présent, le modèle tiendra aussi compte de sa fréquence d'apparition contrairement au modèle de base MIB.
On considère ensuite la loi mixte 2-Poisson (4.3.3) pouvant se réécrire pour les deux cas où le document est jugé pertinent ou non vis-à-vis d'une requête q,

$$\begin{cases} P(TF_j = tf_{t_j,d} | R) := P(E|R) \lambda_1^{tf_{t_j,d}} \frac{e^{-\lambda_1}}{tf_{t_j,d}!} + P(\bar{E}|R) \lambda_2^{tf_{t_j,d}} \frac{e^{-\lambda_2}}{tf_{t_j,d}!} \\ P(TF_j = tf_{t_j,d} | NR) := P(E|NR) \lambda_1^{tf_{t_j,d}} \frac{e^{-\lambda_1}}{tf_{t_j,d}!} + P(\bar{E}|NR) \lambda_2^{tf_{t_j,d}} \frac{e^{-\lambda_2}}{tf_{t_j,d}!} \end{cases} \tag{4.16}$$

Enfin, on va réécrire la mesure RSV^{MIB} du modèle de base MIB (4.2.1) pour obtenir un critère d'ordonnancement propre à BM 25 tenant compte à la fois de la notion d'élitisme et des fréquences des mots dans les documents.

4.4.2 Réécriture du critère d'ordonnancement RSV

Rappelons la formule du critère d'ordonnancement du modèle probabiliste de base MIB (4.2.1) :

$$RSV(d,q) = \sum_{j:t_j=r_j=1} log\left(\frac{p_j(1-s_j)}{s_j(1-p_j)}\right)$$

Notons, les paramètres

$$\begin{cases} \alpha_j = P(E|R) \\ \beta_j = P(E|NR) \end{cases} \tag{4.17}$$

Alors, le critère RSV du modèle BM 25 comporte 4 paramètres à estimer et sa formulation est définie par

Definition 4.4.1 — Critère RSV^{BM25}.

$$RSV^{BM25}(d,q) := \sum_{j:t_j=r_j=1} log\left(\frac{\left(\alpha_j\lambda_1^{tf_{t_j,d}}e^{-\lambda_1} + (1-\alpha_j)\lambda_2^{tf_{t_j,d}}e^{-\lambda_2}\right)\left(\beta_j e^{-\lambda_1} + (1-\beta_j)e^{-\lambda_2}\right)}{\left(\beta_j\lambda_1^{tf_{t_j,d}}e^{-\lambda_1} + (1-\beta_j)\lambda_2^{tf_{t_j,d}}e^{-\lambda_2}\right)\left(\alpha_j e^{-\lambda_1} + (1-\alpha_j)e^{-\lambda_2}\right)}\right) \tag{4.18}$$

avec λ_1, λ_2 les moyennes des lois de Poisson du terme t_j sur les documents élites et non élites puis α_j, β_j les quatre paramètres à estimer du modèle BM 25. L'expression à l'intérieur du logarithme est souvent appelée "poids" et notée δ_j. On appelle également fonction de poids la fonction ayant pour formule celle de RSV^{BM25} et pour paramètre d'entrée δ_j.

Le calcul de cette expression n'est pas facile à ce stade. En 1994, le travail de Roberston et Walker a montré qu'il est possible d'étudier le comportement de l'expression à l'intérieur du logarithme lorsqu'on fait tendre le nombre d'occurrences des termes $tf_{t_j,d}$ vers l'infini.

4.4.3 Approximations du critère RSV^{BM25}

Sous les conditions $\alpha_j > \beta_j$ et $\lambda_1 > \lambda_2$, le travail de Roberston et Walker a montré que pour tout terme t_j lorsque $tf_{t_j,d} \to +\infty$, la mesure RSV^{BM25} tend vers l'approximation $\log \frac{\alpha_j(1-\beta_j)}{\beta_j(1-\alpha_j)}$.

R Nous calculerons cette approximation en exercice dans le TD 3.

Plusieurs approximations du calcul de RSV^{BM25} sont proposées. Toutes ces approximations sont basées sur le fait de chercher à approximer la fonction de poids par une fonction de la même forme, c'est à dire qui doit respecter les caractéristiques suivantes :

- valoir 0 si $tf_{t_j,d} = 0$
- être monotone croissante selon $tf_{t_j,d}$
- avoir un maximum asymptotique
- être approximé par le poids du modèle de base pour un indicateur direct de l'élitisme

De plus, en présence d'une collection de N documents, souvent on choisit les paramètres en considérant que

- sans connaissance a priori sur α_j il est raisonnable de le fixer à 0.5
- $\forall q$ la requête, les documents d'une collection sont en majorité non pertinents, c'est à dire, $\beta_j = \frac{df_{t_j}+0.5}{N}$

Sous ces conditions, pour tout terme t_j, on approxime la mesure du score RSV^{BM25} par

$$RSV^{BM25}(d,q) \approx \sum_{j:t_j=r_j=1} \log \frac{N - df_{t_j} + 0.5}{df_{t_j} + 0.5}. \tag{4.19}$$

R On peut y voir une version lissée des idf des termes idf_{t_j}.

Le travail de Roberston et Walker (1994) a aussi introduit deux autre modifications au calcul de la mesure du score RSV^{BM25}.

La première consiste à tenir compte du nombre d'occurrences normalisé des termes de la requête **dans les documents**. Cela se traduit par la multiplication des poids $\log \delta_j$ par l'expression suivante :

$$\frac{(k_1+1) \times tf_{t_i,d}}{k_1((1-b) + b\frac{L_d}{m}) + tf_{t_i,d}} \tag{4.20}$$

où,

- L_d est la longueur du document d
- $m = \frac{1}{N}\sum_{d \in \mathscr{C}} L_d$ est la moyenne des tailles des documents de la collection
- k_1 est le paramètre contrôlant la prise en compte de la fréquence des termes, par défaut $k_1 = 1.2$
- b est le paramètre contrôlant la prise en compte de la longueur, par défaut $b = 0.75$

R Avec $k_1 = 0$ on retrouve le modèle de base MIB.
Avec $b = 0$ on ne tient plus compte de la longueur de d.

La seconde consiste à tenir compte du nombre d'occurrences normalisé des termes de la requête q **dans la requête elle-même** par la multiplication des poids $\log \delta_j$ par l'expression suivante :

$$\frac{(k_3+1) \times tf_{t_i,q}}{k_3 + tf_{t_i,q}} \tag{4.21}$$

où, k_3 est le paramètre contrôlant la prise en compte de la fréquence $tf_{t_i,q}$, par défaut on fixe $k_3 = 1000$.

En tenant compte de l'ensemble de ces modifications et approximations, on obtient la réécriture finale du critère RSV^{BM25}.

Definition 4.4.2 — Approximation du critère RSV^{BM25}. Le calcul de la mesure du score d'un document d vis-à-vis d'une requête q avec le modèle BM 25 s'écrit selon la formulation suivante :

$$RSV^{BM25}(d,q) = \sum_{j:t_j=r_j=1} \frac{(k_1+1) \times tf_{t_j,d}}{k_1((1-b) + b\frac{L_d}{m}) + tf_{t_j,d}} \times \frac{(k_3+1) \times tf_{t_j,q}}{k_3 + tf_{t_j,q}} \times \log \frac{N - df_{t_j} + 0.5}{df_{t_j} + 0.5} \tag{4.22}$$

où, on rappelle que df_{t_j} est le nombre de documents de la collection contenant le terme t_j

Le modèle BM25 est un des modèles les plus important dans le domaine de la Recherche d'Information que ce soit sur le plan théorique comme sur le plan de la performance. La famille de modèles BM25 est souvent utilisée par les moteurs de recherche même actuels : *Qwant* en est un exemple.

4.4.4 Interprétation de quelques paramètres du critère RSV^{BM25}

Interprétation du paramètre k_1

- k_1 permet de booster le classement de certains documents dont leur nature les auraient discriminée par l'utilisation du score.
- Quels sont ces documents ? Ces documents sont longs et diversifiés. Par exemple, les livres où il est très probable que de nombreux termes différents apparaissent plusieurs fois dans l'œuvre, même lorsque le terme n'est pas le sujet principal du document. On dit que les termes sont « saturés ». La fréquence du terme perd en pouvoir informatif sur la pertinence du document pour ce terme.
- Pourquoi k_1 permet de booster le classement de ces documents ? Car en prenant k_1 grand on peut ainsi faire « grossir » la fréquence $tf_{t_j,d}$ et donc faire en sorte que les scores soient plus élevés pour ces documents particuliers qui par leur nature auraient été moins bien classés. Car tous les termes se retrouvent avec des fréquences élevées et donc peu discriminables les unes des autres.
- Remarque : quand les documents d'une collection sont courts, on conseille de prendre k_1 petit. En effet, il est très peu probable qu'une collection de courts tweets, par exemple, ait un terme plusieurs fois sans être fortement liée à ce terme.

Interprétation du paramètre b

- b grand permet de pénaliser les documents aux sujets non pertinents pour la requête.
- Pourquoi ? Car quand un article est long et qu'on prend b grand, le dénominateur du score devient plus grand. Ainsi pour une fréquence $tf_{t_j,d}$ petite (terme non pertinent) le score sera d'autant plus petit avec notre b grand. On obtient donc bien la pénalisation des documents dont le thème est non pertinent (pour le terme t_j dont on a calculé le $tf_{t_j,d}$).
- Pour b, nous pourrions maintenant nous demander : "quand pensons-nous qu'un document est susceptible d'être très long, et quand cela devrait-il nuire à sa pertinence pour un terme ?" Les documents très spécifiques comme les spécifications techniques ou les brevets sont longs pour être plus précis sur un sujet. Leur longueur ne risque pas de nuire à la pertinence et un b inférieur peut être plus approprié. En revanche, les documents qui touchent à plusieurs sujets différents de manière large - articles de presse (un article politique peut toucher à l'économie, aux affaires internationales et à certaines sociétés), critiques d'utilisateurs, etc. - bénéficient souvent du choix d'un plus grand b afin que les sujets non pertinents pour la recherche d'un utilisateur, y compris le spam et autres, soient pénalisés. Cela rejoint le point précédent.

Ressource : `https://ethen8181.github.io/machine-learning/search/bm25_intro.html`

4.5 Conclusion

Les modèles probabilistes utilisent la théorie des probabilités pour modéliser l'incertitude inhérente au processus de recherche. Ils cherchent à estimer la probabilité d'observer des événements liés au document et à la requête. Ces modèles se différencient soit par les événements qu'ils considèrent soit par les lois qu'ils utilisent. Ils peuvent également se distinguer par les hypothèses qu'ils font ou approximations qu'ils tolèrent. Rappelons que sans prise en compte d'information de pertinence (probabilité de pertinence), le poids des termes est IDF. Ensuite, dans le modèle probabiliste de base on utilise ni les fréquences des termes entre documents ni la longueur des documents. Ainsi, les modèles probabilistes présentés dans ce chapitre sont au fur et à mesure de plus en plus sophistiqués, où le niveau zéro est la prise en compte de la notion de probabilité de pertinence. Cependant, on pourrait avoir envie d'aller plus loin et vouloir considérer un retour de pertinence afin d'améliorer le classement en donnant des estimations de probabilités plus justes.

De nos jours, les modèles probabilistes sont toujours utilisés par les moteurs de recherche du web. En particulier, la famille de modèles BM25, riche de ses approximations et modifications, est très souvent utilisée. Bien sûr ces modèles sont couplés avec des outils supplémentaires permettant d'avoir des informations du type "retour de pertinence" afin d'améliorer les performances du moteur de recherche. Cela peut être une information liée à si vous avez cliqué ou pas sur la liste des résultats, si oui avez-vous cliqué au premier, second ou *r*ème résultats ? Votre historique de navigation peut aussi être pris en compte pour améliorer la performance voir même d'autres type de données personnelles telle que votre localisation, etc.