Chapitre 5 : Modèles probabilistes pour la recherche d'information

- Modèle tri probabiliste (BIR et BM25)
- Modèle de Langue

Introduction

- Pourquoi les probabilités ?
  - La RI est un processus incertain et imprécis
    - Imprécision dans l'expression des besoins
    - Incertitude dans la représentation des informations
  - La théorie des probabilités semble adéquate pour prendre en compte cette incertitude et imprécision

Modèle probabiliste

- 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
  - Les événements qu'ils considèrent
    - P(pert/d, q) : probabilité de pertinence de *d* vis à vis de *q*
    - $P(q,d)$
    - $P(q|d)$
    - $P(d|q)$
  - Les distributions (lois) qu'ils utilisent

RI et probabilité

Modèle probabiliste classique

- *BIR*
- *2-Poisson*
- *Tree Depend.*
- *BM25.*
- *DFR*

Modèle inférentiel

- *Inquery*
- *Modèle de croyances*

Modèle de langue

- *Unigram*
- *Ngram*

Plan

- Chapitre 5.1 :
  - Rappel théorie des probabilités
  - Modèle de tri probabiliste (Probabilistic Ranking Principle)
    - Modèle BIR
    - Modèle BM25 (2-Poisson)
- Chapitre 6 :
  - Modèle inférentiel
- Chapitre 5.2
  - Introduction au modèle de langue
  - Modèle de langue et RI

Rappel probabilités

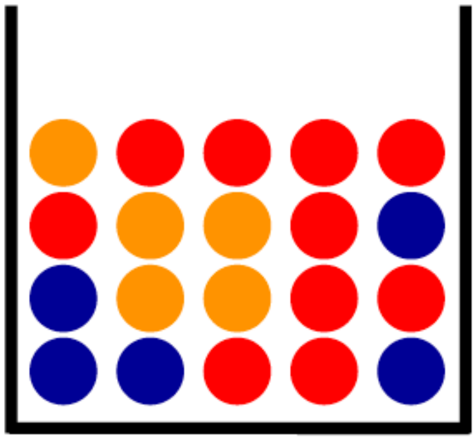
Rappel probabilités

- Probabilité d'un événement
  - P(s) probabilité d'un événement
  - P("pile") = P("face")=1/2
  - Σ P(s) =1 (tous les événements possibles)
  - P(non s) = 1 – P(s)
  - $P(A \cup B)=P(A)+P(B)-P(A \cap B)$

- Probabilité conditionnelle
  - P(s|M) probabilité d'un événement s sachant M
    - P("retrieval" | "information") > P("retrieval" | "politic")

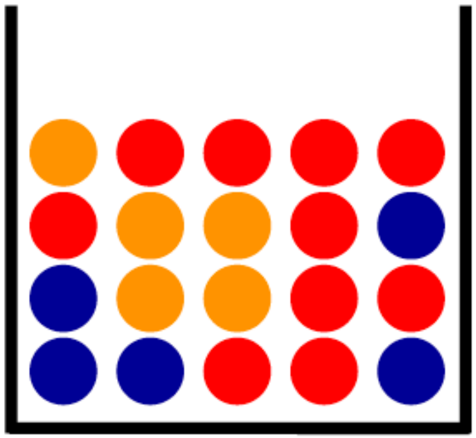
Distribution de probabilités

- Une distribution de probabilités donne la probabilité de chaque évenement
- P(RED) = probabilitié de tirer une boule rouge
- P(BLUE) = probabilité de tirer une boule bleur
- P(ORANGE) =…



Estimation de la distribution de probabilités

- L'estimation de ces probabilités (compter le nombre de cas favorable sur le nombre de cas total)
  - P(Rouge) = ?
  - P(Bleu) = ?
  - P(Orange) = ?
- Deux conditions
  - Probabilité entre 0 et 1
  - La somme des probabilités (de tous les événements est égale à 1)



Rappel probabilités

- Probabilité conditionnelle

$$P(A,B) = P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$ **Règle de Bayes**

– *Evénements indépendants*

  – $P(A,B) = P(A) .P(B)$

– *Evénements dépendants*

  – $P(A,B,C) = P(A) .P(B|A).P(C|A,B)$

– *Formule des probabilités totales*

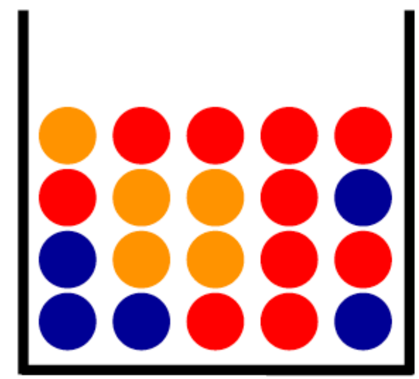
$$P(A) = \sum_i P(A \mid \mathrm{B_i}) * P(\mathrm{B_i})$$

– B1, …Bn est un système complet

Qu'est ce que l'on peut faire avec ces distributions de probabilités

- On peut assigner des probabilités à différentes situations
  - Probabilité de tirer 3 boules orange
  - Probabilité de tirer une orange, une bleue puis une orange

- P() = 0.25
- P() = 0.5
- P() = 0.25 x 0.25 x 0.25
- P() = 0.25 x 0.25 x 0.25
- P() = 0.25 x 0.50 x 0.25
- P() = 0.25 x 0.50 x 0.25 x 0.50



Variable aléatoire

- Une fonction qui associe à chaque résultat d'une expérience aléatoire un nombre (un réel)

  $X : \Omega \rightarrow R$

  $\omega \rightarrow X(\omega)$

- Exemple
  - Jet de deux dés (bleu, rouge), $\Omega=\{(b=1,r=1), (b=1,r=2), \ldots. (b=6,r=6)\}$, la somme S des deux dés est une variable aléatoire discrète à valeurs entre 2 et 12
  - ω est un couple (b, r), $X(\omega) = b+r$ (valeurs possibles, 2, 3, ..12)
  - Ce qui nous intéresse : $P(X=k)$
    - $P(X=2) = 1/36$, $P(X=3)=2/36$, …
- Une VA peut être discrète (ensemble des valeurs est dénombrable) ou continue

Rappel probabilités

- Loi de probabilité d'une variable aléatoire (discrète)
  - Décrit la probabilité de chaque valeur x_i d'une V.A, on note : $p_i = \Pr(X = x_i)$ avec $0 \leq p_i \leq 1$ et $\sum p_i = 1$
  - Loi uniforme : v.a prend ses valeurs X={1,2,…,n}

$$P(X = k) = \frac{1}{n}$$

  - Loi de Bernoulli : X={0,1}

$$P(X = 1) = p$$

$$P(X = 0) = 1 - p$$

$$P(X = x) = p^x (1 - p)^{(1-x)}$$

Rappel probabilités

– Loi binomiale : une v.a. obtenue par le nombre de fois où on a obtenu 1 en répétant *n* fois, indépendamment, une même v.a. de Bernoulli de paramètre *p*, X={0,1, 2, …, n)

$$\Pr(X = k) = \frac{n!}{k!(n-k)!} p^k q^{n-k} \qquad q = 1 - p\,,\ k = 0, \ldots, n$$

– Peut être réécrite

$$\Pr(X_1 = k_1 X_2 = k_2) = \frac{n!}{k_1! k_2!} p_1^{k_1} p_2^{k2} \quad k_1 + k_2 = n \text{ et } p_1 + p_2 = 1$$

– Loi multinomiale : généralisation de la binomiale à m résultats possibles au lieu de 2

$$\Pr(X_1 = k_1 X_2 = k_2 \ldots X_m = k_m) = \frac{n!}{k_1! \ldots k_m!} p_1^{k_1} p_2^{k2} \ldots p_m^{k_n} \qquad \sum_{i=1}^{m} k_i = n$$

Rappel probabilités

– Loi de Poisson :

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

For example, the probability that k =4 sunny days occur in a week; the average is 180/360*7=3. 5 sunny days per week

$$P(k = 4) = 3.4^{4} \frac{.e^{-3.5}}{4!} \approx 0.1888$$

Chapitre 5.1 : Modèle PRP

Probability Ranking Principle (principe du tri probabiliste)

- Probability Ranking Principle (Robertson 1977)
  - "Ranking documents in decreasing order of probability of relevance to the user who submitted the query, where probabilities are estimated using all available evidence, produces the best possible effectiveness"
  - Effectiveness : l'éfficacité est définie en termes de précision

Probability Ranking Principle

- L'idée principale dans PRP
  - Ranking documents in decreasing order of probability of relevance
    - P(pertinent |document) → P(R|d) (ou P(R=1|d))
  - On peut aussi estimer de la même façon la probabilité de non pertinence
    - P(Non pertinence |document) → P(NR|d) (ou P(R=0|d))
  - Un document est sélectionné si : P(R|d)> P(NR|d)
  - Les documents peuvent être triés selon

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

Probabilistic Ranking Principle

- Règle de Bayes

$$P(R \mid d) = \frac{P(d \mid R)P(R)}{P(d)}$$

$$P(NR \mid d) = \frac{P(d \mid NR)P(NR)}{P(d)}$$

- PRP : Ordonner les documents selon

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

$$\stackrel{rank}{=} \frac{P(d \mid R)}{P(d \mid NR)}$$

Comment estimer ces probabilités ?

- Options
  - Comment représenter le document d ?
  - Quelle distribution pour $P(d \mid R)$ et $P(d|NR)$?

- Plusieurs solutions
  - BIR (Binary Independance Retrieval model)
  - "Two poisson model"

Binary Independance Retrieval (BIR)

- Hypothèses
  - 1) Un document est représenté comme un ensemble d'événements (t_i)

$$d = (t_1, \ldots, t_n)$$

    - Un événement t_i dénote la présence ou l'absence d'un terme dans le document

  - 2) Les termes apparaissent dans les documents de manière indépendante

Binary Independence Retrieval (BIR)

- Considérons un document comme une liste de termes

$$RSV(q,d) = \frac{P(d \mid R)}{P(d \mid NR)} = \frac{P((t_1, t_2, \ldots, t_n) \mid R)}{P((t_1, t_2, \ldots, t_n) \mid NR)}$$

- En s'appuyant sur l'hypothèse d'indépendance des termes

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

Binary Independence Retrieval (BIR)

- t_i peut être vu comme, une variable aléatoire (Bernoulli)

  $d=(t_1=x_1\ t_2=x_2\ \ldots\ t_n=x_n)$

$$x_i = \begin{cases} 1 & \textit{terme present} \\ 0 & \textit{terme absent} \end{cases}$$

- $p_i = P(t_i=1|R)$ $\qquad$ $1-p_i = P(t_i=0|R)$
- $q_i = P(t_i=1|NR)$ $\qquad$ $1-q_i = P(t_i=0|NR)$

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

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

Binary Independence Retrieval (BIR)

$$RSV(d,q) = \log \frac{P(d \mid R)}{P(d \mid NR)} = \log \frac{\prod_{i=1}^{n} p_i^{x_i} (1-p_i)^{(1-x_i)}}{\prod_{i=1}^{n} q_i^{x_i} (1-q_i)^{(1-x_i)}}$$

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

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

Constante (quelque soit le document)

Comment estimer p_i and q_i ?

Estimation avec des données d'apprentissage

- En considérant pour chaque terme t_i

Documents	Pertinent (R)	Non-Pertinent (NR)	Total
t_i=1	r	n-r	n
t_i=0	R-r	N-n-R+r	N-n
Total	R	N-R	N

r : nombre de documents pertinents contenant t_i
n: nombre de documents contenant ti
R : nombre total de documents pertinents
N : nombre de documents dans la collection

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

Estimation par maximun de vraisemblance

Estimation des pi et qi

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

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

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

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

Modèle probabiliste BIR

- Lisser les probabilités pour éviter 0

$$RSV(q,d) = \sum \log \frac{\dfrac{r+0.5}{R-r+0.5}}{\dfrac{(n-r+0.5)}{(N-n-R+r+0.5)}}$$

Estimation sans données d'apprentissage

- Estimation de p_i
  - Constante (Croft & Harper 79)
  - Proportionnel à la probabilité d'occurrence du terme dans la collection (n/N)

- Estimation de q_i
  - prendre tous les documents ne comportant pas t_i

$$RSV(Q,D) \overset{Rank}{\approx} \sum_{i=1, d_i=k_i=1}^{k} \log \frac{N - n_i + 0.5}{n_i + 0.5} \qquad \boldsymbol{IDF' = \log \frac{N - n_i}{n_i}}$$

Avantages et inconvénients du modèle BIR

- Avantages
  - Formalisation puissante
  - Modélisation explicite de la notion de pertinence

- Inconvénients
  - La fréquence des termes n'est pas prise en compte
  - Difficulté d'estimer les probabilités sans données d'apprentissage
  - Hypothèse d'indépendance entre termes souvent critiquée …mais pas d'amélioration significative pour les modèles qui considèrent cette dépendance

Modéliser la fréquence des termes

– Point de départ

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

– Hypothèse
  – la v.a t_i prend des valeurs entières x_i qui représentent la fréquence du terme.
  – → Estimer $P(t_i=x_i|R)$: probabilité que t_i apparaisse x_i fois dans les documents pertinents

– Estimation naive
  – Calculer $P(t_i=x_i|R)$ pour tous les x_i potentiels – Px_1, Px_2, Px_3, …
  →plusieurs paramètres à estimer.

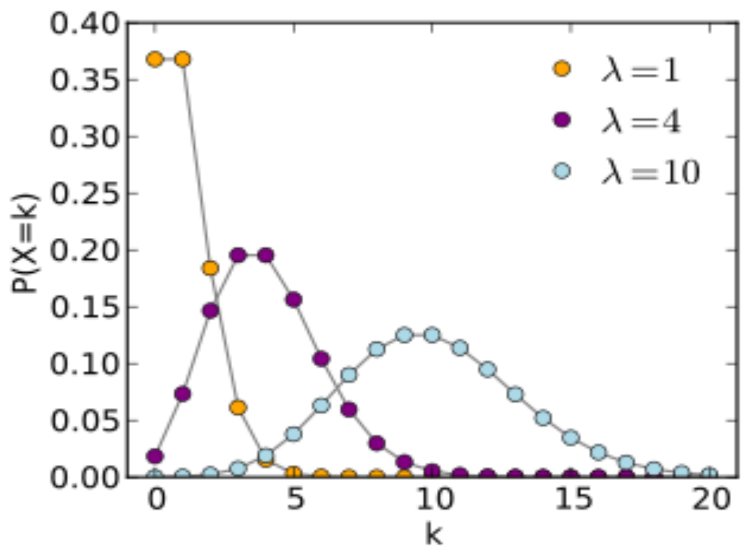
– Modèle paramétrique :
  – On suppose que la v.a t_i suit une certaine loi de probabilité

Modèle 2-Poisson [Harter]

- Idée de base
  - 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

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

    - *λ Moyenne des fréquences des termes dans le document*



Wiki

Modèle 2-Poisson [Harter]

- 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
- On distingue alors
  - Les documents *élites* qui traitent du sujet représenté par le terme
  - Les documents *non élites* qui ne traitent pas du sujet du terme

Modèle 2-Poisson [Harter]

- La distribution des termes dans les documents suit une distribution mixte 2-Poisson

$$P(t = k) = P(E)\lambda_1^k \frac{e^{-\lambda_1}}{k!} + P(\neg E)\lambda_0^k \frac{e^{-\lambda 0}}{k!}$$

  - P(E) : probabilité à priori que le document soit élite
  - λ_1, λ_0 Moyennes des fréquences des termes dans les documents élites et non élites respectivement

Modèle BM25

- Intégrer la notion d'élite dans le calcul des probabilités de pertinence d'un terme
  - $p_i = P(E|R)$
  - $q_i = P(E|NR)$

$$P(t_i = k|\mathrm{R}) = P(E|\mathrm{R})\lambda_1^k \frac{e^{-\lambda_1}}{k!} + P(\neg E|\mathrm{R})\lambda_0^k \frac{e^{-\lambda 0}}{k!}$$

$$P(t_i = k|\mathrm{NR}) = P(E|\mathrm{NR})\lambda_1^k \frac{e^{-\lambda_1}}{k!} + P(\neg E|\mathrm{NR})\lambda_0^k \frac{e^{-\lambda 0}}{k!}$$

Modèle BM25

- Modèle probabiliste de de base

$$P(R|d) \overset{rank}{=} \prod_{i=1}^{n} \frac{P(t_i \in d | R) * P(t_i \notin d | NR)}{P(t_i \in d | NR) * P(t_i \notin d | R)}$$

- Avec les fréquences

$$P(R|d) \overset{rank}{=} \prod_{i=1}^{n} \frac{P(t_i = tf | R) * P(t_i = 0 | NR)}{P(t_i = tf | NR) * P(t_i = 0 | R)}$$

$$P(t_i = tf \mid \mathrm{R}) = p\lambda_1^{tf} \frac{e^{-\lambda_1}}{tf!} + (1-p)\lambda_0^{tf} \frac{e^{-\lambda 0}}{tf!}$$

$$P(t_i = tf \mid \mathrm{NR}) = q\lambda_1^{tf} \frac{e^{-\lambda_1}}{tf!} + (1-q)\lambda_0^{tf} \frac{e^{-\lambda 0}}{tf!}$$

Modèle BM25

- Réécriture de la fonction de tri (en passe au log)

$$P(R \mid d) = \sum_{i=1}^{n} \log\left(\frac{(p\lambda_1^{tf} e^{-\lambda_1} + (1-p)\lambda_0^{tf} e^{-\lambda 0})(qe^{-\lambda_1} + (1-q)e^{-\lambda 0})}{(q\lambda_1^{tf} e^{-\lambda_1} + (1-q)\lambda_0^{tf} e^{-\lambda 0})(pe^{-\lambda_1} + (1-p)e^{-\lambda 0})}\right)$$

- Quatre paramètres à estimer
- S. Walker et S. Robertson ont estimé ces paramètres selon la formule : BM25 (*Roberston et. al sigir 1994*)

Approximation de la fonction de poids

- La fonction de poids doit respecter les caractéristiques suivantes :
  - (a) 0 si *tf*=0
  - (b) monotone croissante avec *tf*
  - (c) a un maximum asymptotique
  - (d) approximé par le poids du modèle de base

Approximation .. (suite)

- On réarrange la fonction

$$P(w) = \log \frac{(p + (1-p)(\lambda_0/\lambda_1)^{tf} e^{\lambda_1 - \lambda_0})(qe^{\lambda_0 - \lambda_1} + (1-q))}{(q + (1-q)(\lambda_0/\lambda_1)^{tf} e^{\lambda_1 - \lambda_0})(pe^{\lambda_0 - \lambda_1} + (1-p))}$$

- $\lambda_0 << \lambda_1 \quad et \quad tf_{\rightarrow \infty}$

$$\rightarrow \qquad P(w) = \log \frac{p(1-q)}{q(1-p)}$$

Approximation.. (suite)

$$P(w) = \frac{(k_1 + 1)tf_i}{K + tf_i} w^1$$

K est une constante

Cette onction respecte les 4 contraintes

La forme (finale) de BM25

$$RSV^{BM25} = \sum_{i \in q} \log \frac{N}{df_i} \cdot \frac{(k_1 + 1)tf_i}{k_1((1-b) + b\frac{dl}{avdl}) + tf_i}$$

- k_1 contrôle term frequency
  - $k_1 = 0$ → modèle BIR;
  - b controle la normalisation de la longueur
  - $b = 0$ → pas de normalisation; $b = 1$ fréquence relative
- k_1 est entre 1.2–2 et b autour de 0.75

Modèle BM25

- BM25 est un des modèles les plus importants dans le domaine de la RI sur les deux plans théorique et performance (rappel précision)