Chapitre 6
Modèle probabiliste
probabilistic model

1

Modèle probabiliste

- 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 de la probabilité semble adéquate pour quantifier (pour mesurer) cette incertitude et imprécision

2

Modèle probabiliste

- Le modèle probabiliste tente d'estimer la probabilité qu'un document donné soit pertinent pour une requête donnée
  - P(pert/d, q) : probabilité de pertinence de *d* vis à vis de *q*
  - *P(q,d)*
  - *P(q/d)*
  - *P(d/q)*

3

Modèle probabiliste de base

- Considérons une requête q et un document d, le modèle probabiliste tente d'estimer la probabilité que le document d appartienne à la classe des documents pertinents (non pertinents)

P(NR|D) P(R|D)

- Un document est sélectionné si : P(R/d) > P(NR/d)

4

Probability Ranking Principle

- Supposons que la pertinence d'un document est indépendante des autres documents
- Probability Ranking Principle (Principe d'appariement probabiliste)
  - "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"
- L'efficacité est définie en termes de précision

5

Probability Ranking principal

- Documents triés selon PRP

RSV(q,d)=O(d)=P(R/d) / P(NR/d)

Rappel Règle de Bayes

$$p(a,b) = p(a \cap b) = p(a \mid b)\,p(b) = p(b \mid a)\,p(a)$$

$$p(\overline{a} \mid b)\,p(b) = p(b \mid \overline{a})\,p(\overline{a})$$

$$p(a \mid b) = \frac{p(b \mid a)\,p(a)}{p(b)}$$

6

Probabilistic Ranking Principle Démonstration

- 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 par rapport

$$RSV\,(q,d) = O(d) \approx \frac{p(d \mid R)}{p(d \,/\, NR)}$$

Probability Ranking principal

- Options :
  - Comment représenter le Document D ?
  - Quelle distribution utilisée pour P(D | R) et P(D|NR)?
  - Etant donnée une requête comment estimer les paramètres du modèle ?
- Plusieurs solutions
  - BIR (Binary Independant Model)
  - "Two poisson model"

Binary Independence Retrieval (BIR)

- Document : ensemble d'événements
- Evénement dénote la présence ou l'absence d'un terme dans un document

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

$t_i = 1$ si un terme est présent dans un document

- **Independence":** les termes apparaissent dans les documents de manière indépendante

L'appariement (Probabilistic Ranking Principle)

- Considérons un document comme une liste de termes
  - P(d/Pert) et P(d/Npert) sont estimées par les probabilités conditionnelles selon qu'un terme de la requête est présent dans un document pertinent ou non prtinent.

$$RSV(q,d) = O(d) \approx \frac{p(d(t_1, t_2, \ldots, t_n) \mid R)}{p(d(t_1, t_2, \ldots, t_n) / NR)}$$

• en se basant sur l'hypothèse d'indépendance

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

Binary Independence Model
Retour sur l'hypothèse d'indépendance

- En probabilité, la combinaison de plusieurs événements doit être déterminée comme suit :

P(t1, t2, t3, t4 ...|R) = P(t1|R) * P(t2|t1,R) * P(t3|t1,t2,R) * P(t4|t1,t2,t3,R) * ...

- En RI, la présence et l'absence de termes sont dépendantes.
  - Par exemple, si le terme «informatique» apparaît dans un document, il y a plus de chance que le terme «ordinateur» apparaît aussi. Ainsi

P(ordinateur=1 | informatique=1) > P(ordinateur=1)

11

Binary Independence Retrieval (BIR)

Loi de Bernoulli

D = $\{t_1=x_1, t_2=x_2, \ldots t_n=x_n\}$

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

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

$$= \prod_{i=1}^{n} P(t_i = 1 \mid R)^{x_i} P(t_i = 0 \mid R)^{(1-x_i)} = \prod_{i=1}^{n} p_i^{x_i} (1-p_i)^{(1-x_i)}$$

$$P(D \mid NR) = \prod_{i=1}^{n} P(t_i = 1 \mid NR)^{x_i} P(t_i = 0 \mid NR)^{(1-x_i)} = \prod_{i=1}^{n} q_i^{x_i} (1-q_i)^{(1-x_i)}$$

12

Binary Independence Retrieval (BIR)

$$Odd(D) = \log \frac{P(D\,|\,R)}{P(D\,|\,NR)} = \log \frac{\prod_{i=1}^{n} p_i^{x_i}(1-p_i)^{(1-x_i)}}{\prod_{i=1}^{n} q_i^{x_i}(1-q_i)^{(1-x_i)}}$$

$$= \sum_{i:x_i=1}^{n} x_i \log \frac{p_i(1-q_i)}{q_i(1-p_i)} + \underbrace{\sum_{i=1}^{n} \log \frac{1-p_i}{1-q_i}}_{\text{Constante (quelque soit le document)}}$$

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

Constante (quelque soit le document)

Comment estimer pi and qi

13

Estimation avec des données d'apprentissage

- En considérant pour chaque terme ti

Documents	Pertinent	Non-Pertinent	Total
ti=1	r	n-r	n
ti=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

14

Estimation par maximun de vraisemblance

Estimation des pi et qi

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

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

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

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

15

Modèle probabiliste BIR

- Lisser les probabilités pour éviter 0,

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

- 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

16

Estimation sans données d'apprentissage

Lorsque des données d'apprentissage pour l'évaluation ne sont pas disponibles

➡ *estimation a priori: on donne des valeurs pour pi et qi*
pi = 0.5
qi = ni / N (l'ensemble des documents non-pertinents est beaucoup plus important que l'ensemble des documents pertinents) revient aussi à considérer qu'on n'a pas d'informations de pertinence dans la formule

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

Pour éviter le zéro, on ajoute 0,5. RSV devient

$$RSV(q,d) = \sum_{i \in (q \cap d)} \log(\frac{(N-ni+0,5)}{(ni+0,5)})$$

17

Pour une Requête et des documents sont pondérés

$$RSV(q,dj) = \sum_{i \in (q \cap d)} Wij * qtfi * \log(\frac{(N-ni+0,5)}{(ni+0,5)})$$

18

Modèle “2-poisson”

- S. Walker et S. Robertson on estimé ces paramètres selon la formule : BM25

$$RSV(q,dj) = \sum_{i \in (q \cap dj)} \left(\frac{(k+1) * tfij}{tfij + k * (1 - b + b * (\frac{dlj}{avgdl}))} \right) * \log\left(\frac{(N - ni + 0{,}5)}{(ni + 0{,}5)}\right)$$

Tfij : la fréquence du terme i dans la document j
ni : le nombre de document contenant le terme i
N: nombre de document de la collection
dlj : longueur de du document j (nombre de termes)
avgdl : longueur moyenne des documents de la collection
k, b : des constantes

19

The BM25 Formula

$$\sum_{T \in Q} w^{(1)} \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf}$$

“Okapi TF/BM25 TF”

here

Q is a query, containing terms T
$w^{(1)}$ is the Robertson/Sparck Jones weight [5] of T in Q

$$\log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} \qquad (2)$$

N is the number of items (documents) in the collection
n is the number of documents containing the term
R is the number of documents known to be relevant to a specific topic
r is the number of relevant documents containing the term
K is $k_1((1-b)+b.dl/avdl)$
k_1, b and k_3 are parameters which depend on the on the nature of the queries and possibly on the database; k_1 and b default to 1.2 and 0.75 respectively, but smaller values of b are sometimes advantageous; in long queries k_3 is often set to 7 or 1000 (effectively infinite)
tf is the frequency of occurrence of the term within a specific document
qtf is the frequency of the term within the topic from which Q was derived
dl and $avdl$ are respectively the document length and average document length measured in some suitable unit.

Modèle probabiliste : récapitulatif

- Un des modèles les plus importants dans le domaine de la RI
- BM25 donne les meilleures performances en terme de rappel et précision
- Problèmes :
  - Indépendance entre les termes
  - Indépendance entre les documents

21