Recherche d'information

Modèles en Recherche d'Information
Modèles probabilistes

Cours Master Recherche Paris 13
Recherche et extraction d'information

Antoine Rozenknop
Source : Romaric Besançon
CEA-LIST/LIC2M

Au menu

- Principe de classement probabiliste
- Modèle probabiliste de base
- Indexation probabiliste
- Modèle Okapi

Modèles probabilistes

- Pourquoi utiliser des modèles probabilistes ?
  - la recherche d'information traite une information incertaine
    - dans la représentation d'un document
    - dans la représentation de la requête
    - dans la mise en correspondance de la requête et du document
    - dans la pertinence du résultat par rapport à la requête
- Les probabilités sont un outil naturel pour essayer de quantifier l'incertitude

Modèles probabilistes

- principe de classement probabiliste (*probability ranking principle*), énoncé dans les années 60/70

> *un système qui retourne pour chaque requête une liste de documents dans l'ordre décroissant de la probabilité que le document est utile à l'utilisateur ayant soumis la requête, en supposant que ces probabilités soient estimées aussi exactement que possible à partir de toute l'information disponible, aura la meilleur performance possible sur la base de cette information.*

- ➔ retourne une liste ordonnée de documents
- ➔ le score d'un document est fonction de sa probabilité de pertinence
- ➔ la pertinence reste une notion floue
- ➔ l'ordre des documents est plus important que leur score

Modèles probabilistes

- Hypothèses simplificatrices du principe de classement probabiliste
  - la pertinence d'un document est indépendante des jugements portés sur les autres documents
  - l'utilité d'un document pertinent ne dépend pas du nombre de documents pertinents que l'utilisateur a déjà obtenus

- Ces hypothèses sont en général également faites dans les autres modèles

Modèle probabiliste de base

- on modélise la pertinence comme un événement probabiliste :
  - pour une requête *Q* donnée, estimer *P(R|D)* la probabilité qu'on obtienne une information pertinente par le document *D*
    - dépend de Q: $P(R|D) = P_Q(R|D)$
  - on peut estimer de la même façon *P(NR|D)* la probabilité de non-pertinence de D
  - on retourne un document *D* si *P(R|D) > P(NR|D)*
  - on donne au document D le poids $s(D) = \frac{P(R|D)}{p(NR|D)}$

Modèle probabiliste de base

- par Bayes

$$P(R|D) = \frac{P(D|R) \cdot P(R)}{P(D)}$$

- P(D|R): probabilité que D fasse partie de l'ensemble des documents pertinents
- P(R): probabilité de la pertinence d'un document quelconque
- P(D): probabilité que D soit choisi

$$P(NR|D) = \frac{P(D|NR) \cdot P(NR)}{P(D)}$$

- P(D|NR): probabilité que D fasse partie de l'ensemble des documents non pertinents
- P(R): probabilité de la non-pertinence d'un document quelconque
- P(D): probabilité que D soit choisi

Modèle probabiliste de base

$$P(R|D) = \frac{P(D|R) \cdot P(R)}{P(D)} \qquad P(NR|D) = \frac{P(D|NR) \cdot P(NR)}{P(D)}$$

- *P(R), P(NR)*, *P(D)* sont constantes
- comme c'est l'ordre qui est important, les modifications de score par des constantes qui ne changent pas l'ordre peuvent être ignorées

➔ le score d'un document dépend donc seulement de *P(D|R)* et *P(D|NR)*

Modèle probabiliste de base

- Dans le modèle probabiliste de base, on suppose une indexation binaire des termes
  ➥ on utilise la présence ou l'absence d'un terme dans les documents pertinents ou non pertinents

- *on note* $\begin{cases} x_i = 1 \ si \ t_i \in D \\ x_i = 0 \ si \ t_i \notin D \end{cases}$

Modèle probabiliste de base

- on suppose l'indépendance des termes, mais on suppose que la distribution des termes dans les documents pertinents ou non pertinents est différente.

$$P(D|R) = \prod_{i=1}^{T} p_i^{x_i} (1-p_i)^{1-x_i} \text{ avec } p_i = P(t_i \in D|R)$$

$$P(D|NR) = \prod_{i=1}^{T} q_i^{x_i} (1-q_i)^{1-x_i} \text{ avec } q_i = P(t_i \in D|NR)$$

Modèle probabiliste de base

- on passe en log

$$\log P(D|R) = \sum_{i=1}^{T} x_i \log p_i + (1-x_i)\log(1-p_i)$$

$$\log P(D|NR) = \sum_{i=1}^{T} x_i \log q_i + (1-x_i)\log(1-q_i)$$

➥ $$\log(s(D)) = \sum_{i=1}^{T} x_i \underbrace{\log \frac{p_i(1-q_i)}{q_i(1-p_i)}}_{w_i} + \underbrace{\sum_{i=1}^{T} \log \frac{1-p_i}{1-q_i}}_{\text{constante}}$$

➥ $$score(D) = \sum_{i=1}^{T} x_i \cdot w_i$$

➥ *on a une formule qui ressemble à un produit scalaire d'un facteur fréquentiel binaire et d'un poids dépendant du terme*

Modèle probabiliste de base

- reste à estimer p_i et q_i

	docs contenant t_i	*docs ne contenant pas t_i*	
docs pertinents	r_i	$n - r_i$	n
docs non pertinents	$R_i - r_i$	$N - R_i - n + r_i$	$N - n$
	R_i	$N - R_i$	N

$$p_i = P(t_i \in D | R) = \frac{r_i}{n} \qquad 1 - p_i = P(t_i \notin D | R) = \frac{n - r_i}{n}$$

$$q_i = P(t_i \in D | NR) = \frac{R_i - r_i}{N - n} \qquad 1 - q_i = P(t_i \notin D | NR) = \frac{N - R_i - n + r_i}{N - n}$$

Modèle probabiliste de base

- on a alors

$$score(D) = \sum_{t_i} x_i \cdot w_i$$

avec

$$w_i = \log \frac{p_i(1-q_i)}{q_i(1-p_i)} = \log \frac{\frac{r_i}{n-r_i}}{\frac{R_i - r_i}{N - R_i - n + r_i}}$$

Modèle probabiliste de base

- pour éviter les 0, un lissage de cette formule est proposé :

$$w_i = \log \frac{\frac{r_i + 0.5}{n - r_i + 0.5}}{\frac{R_i - r_i + 0.5}{N - R_i - n + r_i + 0.5}}$$

Modèle probabiliste de base

- lorsque des données d'apprentissage pour l'évaluation ne sont pas disponibles
  ➥ estimation a priori: on donne des valeurs pour p_i et q_i
  - $p_i = 0.5$
  - $q_i = R_i / N$ (l'ensemble des documents non-pertinents est beaucoup plus important que l'ensemble des documents pertinents)

    ➥ $w_i = \log \frac{N - R_i}{R_i}$ ➡ *on retrouve le facteur idf probabiliste intégré dans le modèle vectoriel*

- revient aussi à considérer qu'on n'a pas d'informations de pertinence dans la formule précédente (*$n=r_i=0$)*

Indexation probabiliste

- Dans un modèle probabiliste, le choix de retenir ou non un terme d'indexation doit être lié à la probabilité qu'un utilisateur désirant ce document écrive ce terme dans la requête.
- ➔ modèle basé sur la distribution théorique des mots dans un document
- ➔ les occurrences d'un mot dans un document sont distribués de façon aléatoire: la probabilité qu'un mot apparaisse k fois dans un document suit une loi de Poisson

$$P(freq(t, D) = k) = \frac{e^{-\lambda} \cdot \lambda k}{k!}$$

λ est la moyenne des occurrences du mot dans un document

Indexation probabiliste

- On s'est rendu compte que la loi de Poisson décrit bien les mots peu ou pas porteurs de sens → distribution aléatoire
- Les mots porteurs de sens ont tendance à apparaître en groupes
- ➔ Les mots porteurs de sens sont ceux dont la distribution s'éloigne de la distribution de Poisson (test du χ^2).

Notion d'élitisme

- Pour essayer d'être plus précis, on veut sélectionner un terme *t* pour représenter un document si ce terme apparaît plus fréquemment dans ce document que dans un autre choisi au hasard.

  ➥ on veut distinguer les distributions des termes dans les documents où ce terme est représentatif et dans ceux où il ne l'est pas

Notion d'élitisme

- notion d'*élitisme*: on distribue l'ensemble des documents entre deux groupes:
  - ceux qui traitent du thème représenté par le terme *t* (dans lesquels le terme *t* sera plus fréquent): ensemble *élite* (*elite set*), noté *E*
  - ceux qui ne traite pas du thème *t*, dans lesquels l'apparition de *t* est marginale
- les distributions du terme *t* dans les deux groupes sont différentes

Distribution des termes

- distribution mixte 2-Poisson

$$P(freq(t,D)=k)=P(freq(t,D)=k|D\in E)P(D\in E)+ P(freq(t,D)=k|D\notin E)P(D\notin E)$$

$$\hookrightarrow \quad P(freq(t,D)=k)=\pi\cdot\frac{e^{-\lambda_1}\cdot\lambda_1^k}{k!}+(1-\pi)\cdot\frac{e^{-\lambda_2}\cdot\lambda_2^k}{k!}$$

  - π est la probabilité *a priori* qu'un document soit dans le groupe élite
  - λ_1 et λ_2 les paramètres des lois de Poisson de distributions des termes à l'intérieur de chaque groupe ($\lambda_1 \geq \lambda_2$)

Indexation probabiliste

- Pour des mots vides, la notion d'élitisme n'a pas lieu d'être : la distribution des mots vides sur tous les documents est la même
- On estime les paramètres λ_1 et λ_2 :
  - si ces estimations donnent des valeurs proches, on est en présence d'un terme peu spécifique
  - si ces estimations donnent des valeurs très différentes, on est en présence d'un terme spécifique qu'il faut utiliser pour décrire les documents élites de ce terme

Indexation probabiliste

- On mesure le degré de recouvrement des deux lois de Poisson par

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

- On décide qu'un terme *t* doit indexer un document *D* si β>0, avec

$$\beta = P(D \in E \,|\, freq(t, D) = k) + z$$

➥ dans ce cas, on utilise β comme pondération associée au terme *t* pour le document *D*.

Indexation probabiliste

- la probabilité qu'un document *D* soit dans l'ensemble *élite* d'un terme sachant le nombre d'occurrence *k* de ce terme dans le document est estimée en utilisant le théorème de Bayes

$$P(D \in E \mid freq(t, D) = k) = \frac{P(freq(t, D) = k \mid D \in E) P(D \in E)}{P(freq(t, D) = k)}$$

$$\hookrightarrow \quad P(D \in E \mid freq(t, D) = k) = \frac{\pi \cdot e^{-\lambda_1} \cdot \lambda_1^k}{\pi \cdot e^{-\lambda_2} \cdot \lambda_1^k + (1-\pi) \cdot e^{-\lambda_2} \cdot \lambda_2^k}$$

Modèle Okapi

- part du modèle probabiliste de base
- tente de prendre en compte les fréquences des mots dans les documents

modèle de base: $$w_i = \log \frac{P(t_i \in D | R) P(t_i \notin D | NR)}{P(t_i \in D | NR) P(t_i \notin D | R)}$$

➔ avec les *tf* : $$w_i = \log \frac{P(freq(t_i, D) = tf | R) P(t_i \notin D | NR)}{P(freq(t_i, D) = tf | NR) P(t_i \notin D | R)}$$

Modèle Okapi

- utilise *l'élitisme* des termes: distribution mixte 2-poisson des fréquences des termes sur les documents
- hypothèse d'indépendance sur l'élitisme : *la fréquence d'un terme dans un document ne dépend que de l'appartenance du document à l'ensemble élite.*

$$P(freq(t_i, D) = tf \mid R) = P(freq(t_i, D) = tf \mid D \in E(t_i))\ P(D \in E(t_i) \mid R)$$
$$+ P(freq(t, D) = tf \mid D \notin E(t_i))\ P(D \notin E(t_i) \mid R)$$

Modèle Okapi

- l'équation précédente devient alors:

$$w_i = \log \frac{(p_i' \lambda_1^{tf} e^{-\lambda_1} + (1-p_i')\lambda_2^{tf} e^{-\lambda_2})(q_i' e^{-\lambda_1} + (1-q_i')e^{-\lambda_2})}{(q_i' \lambda_1^{tf} e^{-\lambda_1} + (1-q_i')\lambda_2^{tf} e^{-\lambda_2})(p_i' e^{-\lambda_1} + (1-p_i')e^{-\lambda_2})}$$

avec $p_i' = P(D \in E(t_i) | R)$ $q_i' = P(D \in E(t_i) | NR)$

λ_1 et λ_2 les moyennes des lois de Poisson de *t* sur les documents élites et non-élites

Modèle Okapi

- estimation de 4 paramètres par terme pour lesquels aucune observation directe n'est possible (l'élitisme est une variable cachée du modèle)

- on cherche à approximer cette fonction par une fonction de même forme:
  - ✔ 0 si *tf*=0
  - ✔ monotone croissante avec *tf*
  - ✔ a un maximum asymptotique
  - ✔ approximé par le poids du modèle de base pour un indicateur direct de l'élitisme

Modèle Okapi

- approximation proposée dans le modèle Okapi

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

avec k_1 une constante

w^1 la fonction du modèle de base

Modèle Okapi

- normalisation par la longueur des documents: longueur relative par rapport à la longueur moyenne des documents

$$L = \frac{dl}{avdl} = \frac{length(d)}{\frac{1}{N}\sum_{i=1}^{N} lenght(d_i)}$$

- en pondérant les facteur tf_i par cette longueur, on obtient

$$w_i = \frac{(k_1+1)tf_i}{k_1 \times \frac{dl}{avdl} + tf_i} w^1$$

Modèle Okapi

- introduction d'un paramètre supplémentaire pour ajuster l'impact de la normalisation par la longueur du document

$$L' = (1-b) + b \times \frac{dl}{avdl}$$

➥ formule usuelle du modèle Okapi (aussi appelé *BM25*)

$$w_i = \frac{tf_i (k_1 + 1)}{k_1 \times \left((1-b) + b \frac{dl}{avdl} \right) + tf_i} w^1$$