



RECHERCHE D'INFORMATION & TRAITEMENT AUTOMATIQUE DU LANGAGE

Cours 2 : RI - modèles d'appariement

2022-23

Benjamin Piwowarski / Laure Soulier



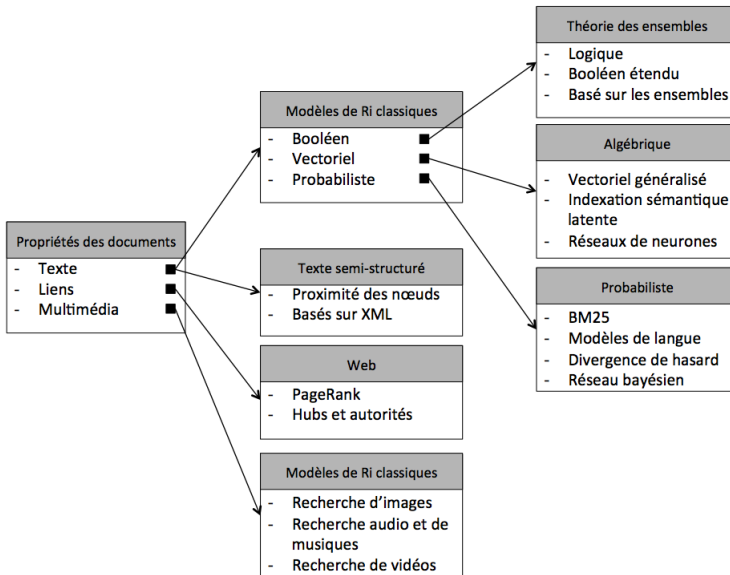
Machine Learning &
Deep Learning for
Information Access

MODÈLES DE RECHERCHE

Hypothèses

- Plus la requête et le document ont de mots en commun, plus grande sera la pertinence du document
- Plus la requête et le document ont une distribution de termes similaire, plus grande sera la pertinence du document

Familles de modèles



MODÈLE BOOLÉEN

Modèle booléen (Salton, 1971)

- Modèle pionnier
- Basé sur la théorie des ensembles
- Représentation logique des documents $L(d)$ et des requêtes $L(q)$ en utilisant les opérateurs logiques : OU (\vee), ET (\wedge) et NON (\neg).
- Exemple :
 - $q = t1 \wedge (\neg t2 \vee t5)$
 - $d1(t1,t3,t5); d2(t1,t3,t5); d3(t1,t2,t3,t4)$
- Score de similarité :

$$RSV(q, d) = \begin{cases} 1 & \text{si } L(q) \subset L(d) \\ 0 & \text{sinon} \end{cases} \quad (1)$$

Inconvénients du modèle booléen

- Pas de pondération de l'importance des termes
- Score de similarité binaire
- Pas d'ordonnancement possible entre les documents sélectionnés
- Risque de sélectionner beaucoup (trop) de documents, surtout lorsque la collection de documents est volumineuse
- Requête peut être difficile à formuler par l'utilisateur

Extensions

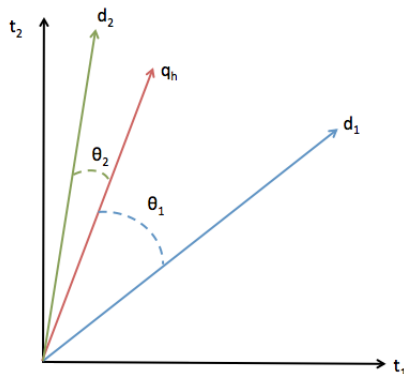
Extensions avec une considération pondérée des poids des termes :

- modèle booléen étendu (Salton and McGill, 1983)
- modèle des ensembles flous (Ogawa et al., 1991)


MODÈLE VECTORIEL

Modèle vectoriel (Salton et al., 1975)

- Espace de caractéristiques t_i , $i = 1 \dots n$ i.e. termes sélectionnés pré-traités
- Représentation des documents - requêtes : vecteur de poids dans l'espace des caractéristiques :
 - document : $d = (x_0, \dots, x_{n-1})$
 - requête : $q = (y_0, \dots, y_{n-1})$



Pondération des vecteurs de représentation

- x_k poids de la caractéristique k dans le document d , e.g. :
 - présence-absence,
 - fréquence du terme dans le document, dans la collection (cf. idf), le plus répandu : $tf*idf$
 - importance du terme pour la recherche
 - facteurs de normalisation (longueur du document)
- Les mots sont supposés indépendants 

Modèle vectoriel : mesures de similarités

- Différentes fonctions de score peuvent être employées avec un codage fréquentiel des documents :

Inner Product	$X \cdot Y$
Mesure de cosinus	$\frac{X \cdot Y}{ X Y }$

Modèle vectoriel : avantages et inconvénients

- **Avantage par rapport au modèle booléen**
 - Les termes sont pondérés
 - Les documents sont évalués sur une échelle continue → Permet la sélection de documents partiellement pertinents
- **Inconvénients**
 - hypothèse d'indépendance des termes + ne tient pas compte de l'ordre des mots ("sac-de-mots" / "bag-of-words")
 - Extension : prise en compte des N-grammes (Song and Croft, 1999)
 - similarité != pertinence. Le document le plus similaire peut-être non pertinent
 - initialement conçu pour des documents courts.
 - Documents longs : facteurs de normalisation, approches hiérarchiques par paragraphes (sélection de paragraphes pertinents + combinaison des scores des paragraphes)

MODÈLE PROBABILISTE

Modèle probabiliste

- Hypothèses et notations

- Espace de probabilité $\mathcal{D} \times \mathcal{Q}$ (document/question)
- Un paire (d, q) est la réalisation d'un tirage aléatoire
- $t \in q$ ($t \in d$) : le terme appartient au document d
- A chaque paire, on associe une variable aléatoire binaire R qui est vraie si le document d est pertinent pour la question q
- $f(q, d) \stackrel{q}{=} g(q, d)$ si f et g ordonnent les documents de la même manière

- Probability Ranking Principle (Robertson 1977)

- Présenter les documents à l'utilisateur selon l'ordre décroissant de leur probabilité de pertinence $P(R|d, q)$
- Propriété : principe optimal car il optimise le risque de Bayes pour la règle de décision suivante : d est pertinent ssi $P(R|d, q) > P(\neg R|d, q)$

Binary independent model

- Score de similarité : rapport des probabilités a posteriori (règle de bayes)

$$\begin{aligned}
 p(R|q, d) &\stackrel{q}{=} \frac{p(R|q, d)}{p(\neg R|q, d)} \\
 &= \frac{p(d|R, q) p(R|q)}{p(d|\neg R, q) p(\neg R|q)} \\
 &\stackrel{q}{=} \frac{p(d|R, q)}{p(d|\neg R, q)}
 \end{aligned}$$

- Note : on pourrait aussi avoir (**autre modèle**)

$$\frac{p(q|R, d) p(R|d)}{p(q|\neg R, d) p(\neg R|d)}$$

Binary independent model

 *Hypothèse* : Les termes apparaissent de manière indépendante

$$p(R|d, q) = \prod_{t \in d} \frac{P(t \in d|R, q)}{P(t \in d|\neg R, q)} \prod_{t \notin d} \frac{P(t \notin d|R, q)}{P(t \notin d|\neg R, q)} \times \frac{P(R|q)}{P(\neg R|q)}$$

avec $p_t = P(t \in d|R, q)$ la probabilité que le terme t apparaisse dans *un document pertinent* pour q , et $u_t = P(t \in d|\neg R, q)$ la probabilité que le terme t apparaisse dans *un document non pertinent* pour q :

$$\begin{aligned} p(R|d, q) &= \frac{P(R|q)}{P(\neg R|q)} \prod_{t \in d} \frac{p_t}{u_t} \prod_{t \notin d} \frac{1 - p_t}{1 - u_t} \\ &= \frac{P(R|q)}{P(\neg R|q)} \prod_{t \in d} \frac{p_t}{u_t} \frac{1 - u_t}{1 - p_t} \prod_{t \in \mathcal{T}} \frac{1 - p_t}{1 - u_t} \\ &\stackrel{q}{=} \prod_{t \in d} \frac{p_t}{u_t} \frac{1 - u_t}{1 - p_t} \end{aligned}$$

Binary independent model

 *Hypothèse* : un terme qui n'appartient pas à la question est uniformément réparti dans les documents pertinents et non pertinents, i.e. $p_t = u_t$

$$p(R|d, q) \stackrel{q}{=} \prod_{t \in d \cap q} \frac{p_t}{u_t} \times \frac{1 - u_t}{1 - p_t}$$

On obtient le score de pertinence :

$$p(R|d, q) \stackrel{q}{=} s(q, d) = \sum_{t \in d \cap q} \log \frac{p_t}{u_t} \times \frac{1 - u_t}{1 - p_t} \quad (2)$$

BIM : Estimation par vraisemblance

✚ Estimation des probabilités p_t et u_t

- Maximum de vraisemblance sur une base d'apprentissage (i.e., fréquences relatives)
- Tableau des fréquences

	Pertinent	Non Pertinent	Total
terme $t \in d$	r_t	$n_t - r_t$	n_t
terme $t \notin d$	$R - r_t$	$N - n_t - R + r_t$	$N - n_t$
total	R	$N - R$	N

- * r_t = nombre de documents pertinents contenant terme t
- * Avec ces fréquences :

$$p_t = \frac{r_t}{R} \text{ et } u_t = \frac{n_t - r_t}{N - R} \quad (3)$$

- * En pratique, on lisse ces fréquences pour éviter les 0 (facteur β : $r+\beta$, etc. . .)

BIM : Estimation par vraisemblance

Formule finale

Que vaut le score de similarité lorsqu'on remplace p_t et u_t par les valeurs de fréquences ?

$$s(q, d) = \sum_{t \in q \cap d} \log \frac{p_t(1 - u_t)}{u_t(1 - p_t)} \quad (4)$$

$$s(q, d) = \sum_{t \in q \cap d} \log \frac{r_t + 0.5}{R - r_t + 0.5} \times \frac{N - n_t - R + r_t + 0.5}{n_t - r_t + 0.5}$$

En pratique, en supposant que $r_t = R = 0$,

$$s(q, d) \approx \sum_{t \in q \cap d} \log \frac{N - n_t + 0.5}{n_t + 0.5} = \sum_{t \in q \cap d} w_t^{(RSJ)} = \sum_{t \in q \cap d} \log \frac{N + 1}{n_t + 0.5} - 1$$

BIM : Estimation par vraisemblance

Formule finale

Que vaut le score de similarité lorsqu'on remplace p_t et u_t par les valeurs de fréquences ?

$$s(q, d) = \sum_{t \in q \cap d} \log \frac{p_t(1 - u_t)}{u_t(1 - p_t)} \quad (4)$$

$$s(q, d) = \sum_{t \in q \cap d} \log \frac{r_t + 0.5}{R - r_t + 0.5} \times \frac{N - n_t - R + r_t + 0.5}{n_t - r_t + 0.5}$$

En pratique, en supposant que $r_t = R = 0$,

$$s(q, d) \approx \sum_{t \in q \cap d} \log \frac{N - n_t + 0.5}{n_t + 0.5} = \sum_{t \in q \cap d} w_t^{(RSJ)} = \sum_{t \in q \cap d} \log \frac{N + 1}{n_t + 0.5} - 1$$

❓ Ça ne vous rappelle rien ?

Modèle probabilistes

Nombreuses variantes / extensions

- longueur des documents (hypothèse implicite d'égale longueur)
- expansion des requêtes
- # doc pertinents considérés (e.g. cas recherche on line \neq off line)
- cooccurrence de termes, prise en compte de "phrases" ...

Okapi - Robertson et al. (1994)

Modèle de référence en RI qui étend BIM

❓ Comment faire pour prendre en compte la fréquence des termes ?

Okapi - Robertson et al. (1994)

Modèle de référence en RI qui étend BIM

❓ Comment faire pour prendre en compte la fréquence des termes ?

💡 Trois idées

- Notion de terme “représentatif” (*elite*) d'un document (et BIM !)
- La fréquence d'un terme dans un document est conditionné par le fait que le terme est *elite* ou non
- La fréquence d'un terme est expliqué par un modèle “2-Poisson” (Harper, 1975)

Okapi - Robertson et al. (1994)

- Notion de groupe élite E_t : plus un terme t apparaît, plus il représente le document
 - * Groupes non élite : modélisation par une loi de Poisson de paramètre μ_t
 - * Groupes élite : Modélisation par une loi de Poisson de paramètre $\lambda_t > \mu_t$
- On suppose que la pertinence R influe directement sur E_t
- On peut donc exprimer la distribution de tf_t d'un terme t de la façon suivante :

$$\begin{aligned}
 P(tf_t|R, q) &= P(tf_t|E_t)P(E_t|R, q) + P(tf_t|\neg E_t)P(\neg E_t|R, q) \\
 &= \mathcal{P}_{\lambda_t}(tf_t) \underbrace{P(E_t|R, q)}_{e_t^1} + \mathcal{P}_{\mu_t}(tf_t) \underbrace{P(\neg E_t|R, q)}_{1-e_t^1}
 \end{aligned}$$

- de même pour $P(tf_t|\neg R, q)$

Okapi - Roberston et al. (1994)

👉 On repart de la formule initiale du modèle BIM en introduisant les fréquences

$$p(R|d, q) = \prod_{t \in d} \frac{P(tf_t|R, q)}{P(tf_t|\neg R, q)} \prod_{t \notin d} \frac{P(TF_t = 0|R, q)}{P(TF_t = 0|\neg R, q)} \times \frac{P(R|q)}{P(\neg R|q)}$$

Okapi - Roberston et al. (1994)

👉 On repart de la formule initiale du modèle BIM en introduisant les fréquences

$$\begin{aligned}
 p(R|d, q) &= \prod_{t \in d} \frac{P(tf_t|R, q)}{P(tf_t|\neg R, q)} \prod_{t \notin d} \frac{P(TF_t = 0|R, q)}{P(TF_t = 0|\neg R, q)} \times \frac{P(R|q)}{P(\neg R|q)} \\
 &\stackrel{q}{=} \sum_{t \notin d \cap q} \frac{P(tf_t|R, q)}{P(tf_t|\neg R, q)} \times \frac{P(TF_t = 0|\neg R, q)}{P(TF_t = 0|R, q)}
 \end{aligned}$$

Okapi - Roberston et al. (1994)

👉 On repart de la formule initiale du modèle BIM en introduisant les fréquences

$$\begin{aligned}
 p(R|d, q) &= \prod_{t \in d} \frac{P(tf_t|R, q)}{P(tf_t|\neg R, q)} \prod_{t \notin d} \frac{P(TF_t = 0|R, q)}{P(TF_t = 0|\neg R, q)} \times \frac{P(R|q)}{P(\neg R|q)} \\
 &\stackrel{q}{=} \sum_{t \notin d \cap q} \frac{P(tf_t|R, q)}{P(tf_t|\neg R, q)} \times \frac{P(TF_t = 0|\neg R, q)}{P(TF_t = 0|R, q)} \\
 &\stackrel{q}{=} \sum_{t \notin d \cap q} \log \frac{\mathcal{P}_{\lambda_t}(tf_t) e_t^1 + \mathcal{P}_{\mu_t}(tf_t) (1 - e_t^1)}{\mathcal{P}_{\lambda_t}(tf_t) e_t^0 + \mathcal{P}_{\mu_t}(tf_t) (1 - e_t^0)} \\
 &\quad + \log \frac{\mathcal{P}_{\lambda_t}(0) e_t^0 + \mathcal{P}_{\mu_t}(0) (1 - e_t^0)}{\mathcal{P}_{\lambda_t}(0) e_t^1 + \mathcal{P}_{\mu_t}(0) (1 - e_t^1)} \\
 &= \sum_{t \notin d \cap q} w_t^{(BM25)}
 \end{aligned}$$

Okapi - Roberston et al. (1994)

❓ Impossible d'estimer λ_t et μ_t facilement...

Okapi - Roberston et al. (1994)

❓ Impossible d'estimer λ_t et μ_t facilement...

💡 On regarde le comportement aux limites

- Pour tf_t à zéro

$$tf_t = 0 \implies w_t = 0$$

- Pour tf_t très grand (on a $\frac{\mathcal{P}_{\lambda}(k)}{\mathcal{P}_{\mu}(k)} = e^{\mu-\lambda} \left(\frac{\lambda}{\mu}\right)^k$)

$$\lim_{tf_t \rightarrow \infty} \frac{e_t^1 + \frac{\mathcal{P}_{\mu_t}(tf_t)}{\mathcal{P}_{\lambda_t}(tf_t)}(1 - e_t^1)}{e_t^0 + \frac{\mathcal{P}_{\mu_t}(tf_t)}{\mathcal{P}_{\lambda_t}(tf_t)}(1 - e_t^0)} = \frac{e_t^1}{e_t^0}$$

$$\frac{\frac{\mathcal{P}_{\lambda_t}(0)}{\mathcal{P}_{\mu_t}(0)} e_t^0 + (1 - e_t^0)}{\frac{\mathcal{P}_{\lambda_t}(0)}{\mathcal{P}_{\mu_t}(0)} e_t^1 + (1 - e_t^1)} \approx \frac{1 - e_t^0}{1 - e_t^1}$$

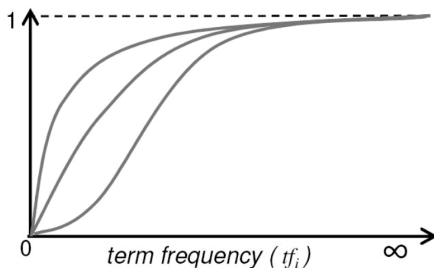
- ... ce qui donne (en revenant à BIM $t \in d \Leftrightarrow E_t$)

$$\lim_{tf_t \rightarrow \infty} w_t^{(BM25)} \approx \log \frac{e_t^1}{e_t^0} \times \frac{1 - e_t^0}{1 - e_t^1} = \log \frac{P(E_t|R, q)}{P(E_t|\neg R, q)} \times \frac{P(\neg E_t|\neg R, q)}{P(\neg E_t|R, q)} = w_t^{(RSJ)}$$

Okapi - Roberston et al. (1994)

❓ Impossible d'estimer λ_t et μ_t facilement...

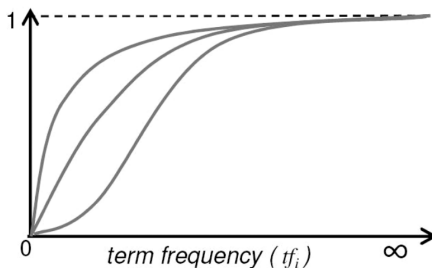
💡 Bon là c'est plus... heuristique : en faisant varier λ_t et μ_t , on obtient ces graphes



Okapi - Roberston et al. (1994)

❓ Impossible d'estimer λ_t et μ_t facilement...

💡 Bon là c'est plus... heuristique : en faisant varier λ_t et μ_t , on obtient ces graphes

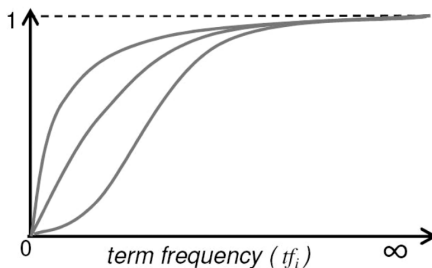


ça ressemble à une fonction de type $\frac{tf_t}{tf_t + \text{constante}}$

Okapi - Roberston et al. (1994)

❓ Impossible d'estimer λ_t et μ_t facilement...

💡 Bon là c'est plus... heuristique : en faisant varier λ_t et μ_t , on obtient ces graphes



ça ressemble à une fonction de type $\frac{tf_t}{tf_t + \text{constante}}$

en mettant tout bout à bout on a

$$w_t^{(BM25)} = w_t^{(RSJ)} \times \frac{tf_t}{tf_t + \text{constante}}$$

Okapi - Roberston et al. (1994)

👉 Dernière étape : Prise en compte de la longueur

Texte verbeux ($b = 1$)

un terme se répète parce que le document est long

Texte multi-thématique ($b = 0$)

un terme se répète parce qu'il est important

💡 Pour régler entre les deux, on utilise B pour normaliser la fréquence d'un terme

$$tf'_t = \frac{tf_t}{B} \text{ avec } B = \left((1 - b) + b \frac{\text{longueur de } d}{\text{longueur moyenne}} \right)$$

ce qui nous donne

$$w_t^{(BM25)} = \frac{tf_t}{tf_t + k_1 \times \left((1 - b) + b \frac{\text{longueur de } d}{\text{longueur moyenne}} \right)} \times \underbrace{\log \frac{N - n_t + 0.5}{n_t + 0.5}}_{w_t^{(RSJ)}}$$

avec k_1 et b constantes (par défaut resp 1.2 et 0.75)

MODÈLES DE LANGUE

Modèles de langue (Ponte, Croft, Hiemstra, ... 98-99)

- Intuition :
 - Modélise la distribution des mots dans une langue
 - Mesure la probabilité d'observer une séquence de mots dans une langue
 - Identifie la source qui a permis de générer un texte

- Formalisation

$$\text{Score}(d, q) = P(s|\theta_M) \quad (5)$$

- Taille des séquences ?
- Estimer la probabilité de chaque séquence ?
- Estimer le modèle de langue ?

Taille des séquences et probabilités

$$P(\bullet \bullet \bullet \bullet) \\ = P(\bullet) P(\bullet | \bullet) P(\bullet | \bullet \bullet) P(\bullet | \bullet \bullet \bullet)$$

- Unigram Models (Assume word independence)

$$P(\bullet) P(\bullet) P(\bullet) P(\bullet)$$

- Bigram Models

$$P(\bullet) P(\bullet | \bullet) P(\bullet | \bullet \bullet) P(\bullet | \bullet \bullet \bullet)$$

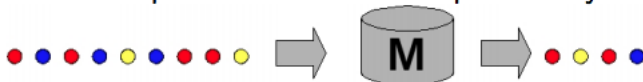
- There are others ...

Modèle de langue du document

- Usually we do not know the model **M**, but have a sample representative of that model

$$P(\text{● ● ● ●} \mid M(\text{● ● ● ● ● ● ● ●}))$$

- First estimate a model from a sample
- Then compute the observation probability



Modèles de langue (Ponte, Croft, Hiemstra, ... 98-99)

- Modèle de langue multinomial
 - Indépendance des termes dans le document
 - Pas d'ordre

$$p(t_1, \dots, t_n) = \prod_t P(TF(t) = tf(t) | \theta_{Md})^{tf(t)} \stackrel{q}{=} \sum_t tf(t) \log(P(t | \theta_{Md})) \quad (6)$$

- Comment estimer ces probabilités ?

$$\operatorname{argmax}_{p_t} \sum_t tf(t) \log(p_t) \quad (7)$$

→ Solution avec le Lagrangien :

$$p_t = \frac{tf(t)}{\sum_t tf(t)} \quad (8)$$

Modèles de langue (Ponte, Croft, Hiemstra, ... 98-99)

- Dans le cas où un mot de la requête n'apparaît pas dans le document d
 - Score du document est égal à 0
 - En pratique, on utilise un lissage de cette probabilité : modèle de mélange multinomial entre la distribution des termes dans le document et la distribution des termes dans la collection
 - * Jelinek-Mercer ($\lambda = 0.8$ pour les requêtes courtes et 0.2 pour les requêtes longues)

$$P(t|d) = (1 - \lambda_t)P(t|\theta_{M_d}) + \lambda_t P(t|\theta_{M_C}) \quad (9)$$

- * Dirichlet

$$\frac{tf(t, d) + \mu p(t|\theta_C)}{length(d) + \mu} \quad (10)$$

Probabilités et Document Prior

- Rappel : $P(d|q) = P(q|d)P(d)/P(q) \propto P(q|d)P(d)$
- $P(d)$ est généralement considéré comme uniforme $\rightarrow P(d|q) \propto P(q|d)$
- $P(d)$ permet également d'intégrer des connaissances a priori dans le calcul de la probabilité :
 - Longueur du document
 - Longueur moyenne des mots
 - Date de publication : "fraîcheur"
 - Nombre de liens
 - PageRank
 - ...

REFORMULATION DE REQUÊTES

Reformulation de requêtes

- Intuition

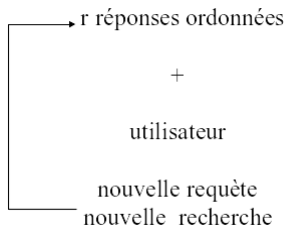
- Difficile de formuler les requêtes qui correspondent aux documents de la collection
 - * On ne sait pas forcément exprimer ce que l'on cherche
 - * On ne sait pas forcément à quoi ressemble le document
- La première requête est souvent naïve, permettant d'avoir une première idée de ce que l'on peut trouver. On peut alors reformuler la requête à partir des résultats :
 - * Etendre la requête originale avec des nouveaux termes
 - * Re-pondérer la requête (étendue)

Reformulation de requêtes

- Relevance feedback : basée sur les feedbacks des utilisateurs
- Pseudo-relevance feedback : basée sur l'analyse locale des documents retournés
- Analyse globale des documents de la collection

Relevance feedback

- Méthode classique



- relevance : valeurs dans $\{0, 1\}$
- idée : utilisateur examine une partie des meilleurs documents et les étiquette 1/0
- la requête est reformulée (enrichissement)

Relevance feedback

- Liste ordonnée des r meilleurs documents

$$D_r(q) = d_1, \dots, d_r \quad (11)$$

- Partition de ces r documents par l'utilisateur

$$D_r(q) = \{D_r^{rel}(q) \cup D_r^{non-rel}(q)\} \quad (12)$$

- Principe du relevance feedback :

$$q' = f(q, D_r^{rel}(q), D_r^{non-rel}(q)) \quad (13)$$

Relevance feedback - Rocchio 1971

- Modèle de base de l'expansion/reformulation de requêtes

$$\vec{Q} = (a.\vec{Q}_0) + (b.\frac{1}{|D_{rel}|} \sum_{d+ \in D_{rel}} \vec{d+}) - (c.\frac{1}{|D_{non-rel}|} \sum_{d- \in D_{non-rel}} \vec{d-}) \quad (14)$$

- Améliorations allant de 20% à 80% par rapport à sans RF
- Différentes variantes :
 - considérer seulement les documents pertinents / que les non-pertinents
 - optimiser a, b, c
 - optimiser le nombre de documents du feedback

Relevance feedback : limites

- Le feedback des utilisateurs n'est pas toujours fiable
 - Positif/négatif ?
 - Degré de pertinence ?
 - Pourquoi ce document peut être pertinent ?
 - En pratique, cela marche bien car on bénéficie de l'effet de masse ("wisdom of the crowd")

Analyse locale : pseudo-relevance feedback

- Suggestion de requête automatique :
 - Pas besoin du feedback utilisateur : les k premiers documents sont considérés comme pertinents
- Approches :
 - Clustering
 - Similarité des terms
 - Analyse des sessions
- Problèmes
 - le système va fournir des documents similaires à ceux déjà trouvés...
 - “Query drift” : si les top documents ne sont pas pertinents, la requête reformulée ne reflètera jamais le besoin de l'utilisateur
 - Peut s'avérer coûteux en termes d'exécution

Analyse globale

- Principe :
 - Etendre la requête à partir de la collection
 - Pas d'assistance de l'utilisateur

Approche : construire le thésaurus des co-occurrences des termes pour identifier les termes les plus proches de ceux de la requête

A vous de jouer... TD is coming!



References I